理解科学丛书 · 卢昌海科普著作




# 因为星星在那里

## 科学殿堂的砖与瓦

卢昌海◎著

数学的纯粹、物理的绚烂……

**原子里的奥秘、星空中的未知……**

迷人的科学知识、锐利的科学方法……

清華大學出版社

理解科学丛书

# Because Stars are There：
# Bricks and Tiles of the Temple of Science

# 因为星星在那里：科学殿堂的砖与瓦

卢昌海　著

清华大学出版社
北　京

**图书在版编目(CIP)数据**

因为星星在那里：科学殿堂的砖与瓦/卢昌海著.--北京：清华大学出版社，2015
(理解科学丛书)
ISBN 978-7-302-40066-0

Ⅰ. ①因…　Ⅱ. ①卢…　Ⅲ. ①科学知识—青少年读物　Ⅳ. ①Z228.1 ②N49

中国版本图书馆 CIP 数据核字(2015)第 089273 号

**责任编辑**：邹开颜
**封面设计**：
**责任校对**：刘玉霞
**责任印制**：

**出版发行**：清华大学出版社
**网　　址**：http://www.tup.com.cn，http://www.wqbook.com
**地　　址**：北京清华大学学研大厦 A 座　**邮　　编**：100084
**社 总 机**：010-62770175　**邮　　购**：010-62786544
**投稿与读者服务**：010-62776969，c-service@tup.tsinghua.edu.cn
**质量反馈**：010-62772015，zhiliang@tup.tsinghua.edu.cn
**印 刷 者**：
**装 订 者**：
**经　　销**：全国新华书店
**开　　本**：165mm×240mm　**印　张**：17　**字　　数**：242 千字
**版　　次**：2015 年 6 月第 1 版　**印　　次**：2015 年 6 月第1次印刷
**印　　数**：1～　000
**定　　价**：　.00 元

产品编号：057284-01



**谨以本书献给我的家人**

# 序言

BECAUSE STARS ARE THERE:
BRICKS AND TILES OF THE TEMPLE OF SCIENCE

作为一位出版了四本书的作者，如果要用一句话来概括写书的感觉的话，那就是：写书比写文章累。这貌似是一句显而易见的大白话，对我这种在写作上有一定兴趣，甚至以写作为乐的人来说，却是一种只有经历过了才意识到的新感觉。这新感觉的起因也是一句显而易见的大白话，那就是：书比文章长。不过，这个“长”对我来说与其说是篇幅之长，不如说是指所费时间之长。因为在一本书的写作过程中，我得不断约束自己的阅读兴趣，把主要精力投注于单一主题。另一方面，我的写作速度又比较慢（或美其名曰“认真”），从而使得写作过程往往长到了对题材的兴趣将尽而书稿远未完成的程度。这时候，写书就变成了对恒心和毅力的考验，而我——很遗憾地——曾两度在这种考验面前失败过，致使《黎曼猜想漫谈》和《从奇点到虫洞》“烂尾”多年（对这一“丢人”事迹感兴趣的读者可参

阅那两本书的后记)，其“累”亦由此可见。

在这种感觉下，若有谁愿把我的文章汇集成书出版，让我既免除写书之累，又可得出书之乐，那对我来说简直就是“天上掉馅儿饼”的美事，几乎要让我生出一种“偷懒”的愧疚了。最近，这样的美事居然落在了我的头上——清华大学出版社愿意出版我的两本文章合集，一本收录科学史方面的文章，一本收录科普方面的文章。

兴奋之下，我很快选好了篇目，但问题来了：一堆文章汇集在一起，以什么作为书名呢？当然，假如我是著名作者，这根本就不是问题，大可取名为“卢昌海科学史作品集”和“卢昌海科普作品集”。但对于明显不著名的我来说，就算不怕僭越地将自己的名字厚颜纳入书名，也只会成为“票房毒药”，因此必须另谋思路。读者可能会笑话我这么小的事情都不能轻松搞定，其实非独我如此，像阿西莫夫(Isaac Asimov)那样的大牌作家也常常为书名发愁呢，以至于在文章合集 *The Sun Shines Bright* 的简介中感慨地说，他几乎想用数字编号来作书名了——当然，他发愁的原因跟我是不同的，他那是因为作品实在太多，显而易见的书名几乎用遍了。

经过思考，为了让两本书略显对仗，我提议将科学史合集取名为《科学殿堂的人和事》，将科普合集取名为《科学殿堂的砖与瓦》。但编辑看了之后觉得这两个标题太平淡。于是我又绞尽脑汁想了半天，却没再想出什么点子来。无奈之下，我决定效仿阿西莫夫，他虽然也为书名发愁，点子可比我多多了，在 *The Sun Shines Bright* 的简介中做完了用数字编号作为书名的“白日梦”后，随即采用了一个颇有些取巧的办法，那就是从所汇集的文章中选取一篇的标题作为书名。现在您所看到的这两本文章合集的书名——《小楼与大师：科学殿堂的人和事》和《因为星星在那里：科学殿堂的砖与瓦》——便也是如此而来。

关注我文章的读者或许注意到了，收录在这两本书中的某些文章是曾经在杂志或报纸上发表过的。不过，杂志和报纸大都有自己固定的风格，有时不免需要作者“削足适履”来契合之。因此，发表在杂志和报纸上的版本与我自

己的版本相比大都存在一定的缺陷,比如经过编辑的改动,以及因字数所限作过删节等。此外,发表在杂志上的版本大都略去了注释及对人名和术语的英文标注等,这其中后者——即英文标注——或许并不重要,但前者——即注释——其实是颇为重要的,往往起着补充正文、澄清歧义等诸多作用。所有这些缺陷在此次汇集成书时都尽可能予以消除了。

与以前的四本书一样,这两本书也是非常接近原稿风格的,在个别细节上甚至有可能略胜于原稿,因为编辑订正的个别错别字由于未曾标注,我未必能在阅读校样时一一察觉并在自己的版本上做出相应的订正。在尊重原稿这个最至关重要的特点上,我要再次对清华大学出版社表示感谢,感谢其对我作品及写作风格的长期——从出版第一本书至今已五年了,够得上用这个词了吧——信任和支持。

最后,希望读者们喜欢这两本新书。

# 目 录

BECAUSE STARS ARE THERE:
BRICKS AND TILES OF THE TEMPLE OF SCIENCE

第一部分

## 数学

第二部分

# 物理

第三部分

## 星际旅行漫谈

第四部分

## 其他

# 第一部分　数　　学

# 孪生素数猜想[①]

2003 年 3 月 28 日,在美国数学研究所(American Institute of Mathematics)位于加州帕洛阿尔托(Palo Alto)的总部,一群来自世界各地的数学家怀着极大的兴趣聆听了圣荷西州立大学(San José State University)数学教授戈德斯通(Daniel Goldston)所做的一个学术报告。在这个报告中,戈德斯通介绍了他和土耳其海峡大学(Boğaziçi University)的数学家伊尔迪里姆(Cem Yıldırım)在证明孪生素数猜想(twin prime conjecture)方面所取得的一个进展。这一进展——如果得到确认的话——将把人们在这一领域中的研究大大推进一步。

那么,什么是孪生素数(twin prime)? 什么是孪生素数猜想? 戈德斯通和伊尔迪里姆所取得的进展又是什么呢? 本文将对这些问题做一个简单介绍。

要介绍孪生素数,首先当然要说一说素数(prime number)这一概念。素数

① 本文撰写于 2003 年 4 月,是我的第一篇数学科普,填补了作为本人兴趣主要组成部分之一的数学在我网站的空白。自那以后,本文曾以"补注"形式对若干后续进展作了简单提及,并于 2014 年 9 月进行了不改变基本结构的轻微修订。

是除了1和自身以外没有其他因子的自然数。在数论中，素数可以说是最纯粹、也最令人着迷的概念。关于素数，一个最简单的事实就是：除了2以外，所有素数都是奇数(因为否则的话，除了1和自身以外还会有一个因子2，从而不满足定义)。由这一简单事实可以得到一个简单推论，那就是：大于2的两个相邻素数之间的最小可能的间隔是2。所谓孪生素数指的就是这种间隔为2的相邻素数，它们之间的距离已经近得不能再近了，就像孪生兄弟一样。不难验证，在孪生素数中，最小的一对是(3，5)，在100以内则还有(5，7)、(11，13)、(17，19)、(29，31)、(41，43)、(59，61)和(71，73)等另外7对，总计为8对。进一步的验证还表明，随着数字的增大，孪生素数的分布大体上会变得越来越稀疏，寻找孪生素数也会变得越来越困难。

那么，会不会在超过某个界限之后就再也不存在孪生素数了呢？

这个问题让我们联想到素数本身的分布。我们知道，素数本身的分布也是随着数字的增大而越来越稀疏的，因此也有一个会不会在超过某个界限之后就再也不存在的问题。不过幸运的是，早在古希腊时代，著名数学家欧几里得(Euclid)就证明了素数有无穷多个(否则的话——即假如素数没有无穷多个的话——今天的许多数论学家恐怕就得另谋生路了)。长期以来数学家们普遍猜测，孪生素数的情形与素数类似，虽然其分布随着数字的增大而越来越稀疏，总数却是无穷的。这就是与哥德巴赫猜想(Goldbach conjecture)齐名、集令人惊异的表述简单性与令人惊异的证明复杂性于一身的著名猜想——孪生素数猜想。

*孪生素数猜想：存在无穷多个素数 $p$，使得 $p+2$ 也是素数。*

究竟是谁最早明确地提出这一猜想我没有考证过，但1849年法国数学波利尼亚克(Alphonse de Polignac)曾提出过一个猜想：对于任意偶数 $2k$，存在无穷多组以 $2k$ 为间隔的素数。这一猜想被称为波利尼亚克猜想(Polignac's conjecture)。对于 $k=1$，它就是孪生素数猜想。因此人们有时把波利尼亚克作为孪生素数猜想的提出者。值得一提的是，人们对不同的 $k$ 所对应的素数对

的命名是很有趣的：$k=1$(即间隔为 2)的素数对我们已经知道叫做孪生素数；$k=2$(即间隔为 4)的素数对被称为 cousin prime(表兄弟素数)，比“孪生”稍远；而 $k=3$(即间隔为 6)的素数对竟被称为 sexy prime！这回该相信“书中自有颜如玉”了吧？不过别想歪了，之所以称为 sexy prime，其实是因为 sex 正好是拉丁文中的“6”(因此 sexy prime 的中文译名乃是毫无联想余地的“六素数”)。

孪生素数猜想还有一个更强的形式，是英国数学家哈代(Godfrey Hardy)和李特伍德(John Littlewood)于 1923 年提出的，有时被称为哈代-李特伍德猜想(Hardy-Littlewood conjecture)或强孪生素数猜想(strong twin prime conjecture)①。这一猜想不仅提出孪生素数有无穷多组，而且还给出其渐近分布为

$$\pi_2(x) \sim 2C_2\int_2^x \frac{\mathrm{d}t}{(\ln t)^2}$$

其中 $\pi_2(x)$ 表示小于 $x$ 的孪生素数的数目，$C_2$ 被称为孪生素数常数(twin prime constant)，其数值为

$$C_2 = \prod_{p\geqslant 3} \frac{p(p-2)}{(p-1)^2} \approx 0.660\,161\,181\,584\,686\,957\,392\,781\,211\,001\,45\cdots$$

强孪生素数猜想对孪生素数分布的拟合程度可以由表 1 看出。很明显，拟合程度是相当漂亮的。假如可以拿观测科学的例子来作比拟的话，如此漂亮的拟合几乎能跟英国天文学家亚当斯(John Couch Adams)和法国天文学家勒维耶(Urbain Le Verrier)运用天体摄动规律对海王星位置的预言，以及爱因斯坦(Albert Einstein)的广义相对论对光线引力偏转的预言等最精彩的观测科学成就相媲美，可以算同为理性思维的动人篇章。这种拟合对于纯数

① 确切地说，哈代和李特伍德于 1923 年所提出的猜想共有两个，分别称为第一哈代-李特伍德猜想(first Hardy-Littlewood conjecture)和第二哈代-李特伍德猜想(second Hardy-Littlewood conjecture)。其中第一哈代-李特伍德猜想又称为 $k$-tuple 猜想($k$-tuple conjecture)，它给出了所有形如($p, p+2m_1, \cdots, p+2m_k$)(其中 $0<m_1<\cdots<m_k$)的素数 $k$-tuple 的渐近分布。强孪生素数猜想只是 $k$-tuple 猜想的一个特例。

学的证明来说虽起不到实质帮助，却大大增强了人们对孪生素数猜想的信心。

**表 1**

| $x$ | 孪生素数数目 | 强孪生素数猜想给出的数目 |
|---|---|---|
| 100 000 | 1 224 | 1 249 |
| 1 000 000 | 8 169 | 8 248 |
| 10 000 000 | 58 980 | 58 754 |
| 100 000 000 | 440 312 | 440 368 |
| 10 000 000 000 | 27 412 679 | 27 411 417 |

在这里还可以顺便提一下，强孪生素数猜想所给出的孪生素数分布规律可以通过一个简单的定性分析来“得到”①：我们知道，素数定理（prime number theorem）表明对于足够大的 $x$，在 $x$ 附近素数的分布密度大约为 $1/\ln(x)$，因此两个素数位于宽度为 2 的区间之内（即构成孪生素数）的概率大约为 $2/\ln^2(x)$。这几乎正好就是强孪生素数猜想中的被积函数——当然，两者之间还差了一个孪生素数常数 $C_2$，而这个常数显然正是哈代和李特伍德的功力深厚之处②。

除了强孪生素数猜想与孪生素数实际分布之间的漂亮拟合外，对孪生素数猜想的另一类“实验”支持来自于对越来越大的孪生素数的直接寻找。就像对大素数的寻找一样，这种寻找在很大程度上成为了对计算机运算能力的一种检验。1994 年 10 月 30 日，这种寻找竟然使人们发现了英特尔（Intel）奔腾（Pentium）处理器浮点除法运算的一个瑕疵（bug），在工程界引起了不小的震动。截至 2002 年底，人们发现的最大的孪生素数是：

$$(33\,218\,925\times 2^{169\,690}-1,\ 33\,218\,925\times 2^{169\,690}+1)$$

这对素数中的每一个都长达 51 090 位。许多年来这种纪录一直被持续而成功地刷新着，它们对于纯数学的证明来说虽也起不到实质帮助，却同样有助于

① 这种定性分析被澳大利亚数学家陶哲轩（Terence Tao）称为“概率启发式理由”（probabilistic heuristic justification），它不是证明，但对于判断命题成立与否有一定的启示性。

② 对孪生素数常数 $C_2$ 也存在“概率启发式理由”，感兴趣的读者可参阅美国数学家查基尔（Don Zagier）的“The First 50 Million Prime Numbers”，Math. Intl. 0，221-224（1977）。

增强人们对孪生素数猜想的信心①。

好了，介绍了这么多关于孪生素数的资料，现在该说说人们在证明孪生素数猜想上所走过的征途了。

迄今为止，在证明孪生素数猜想上的成果大体可以分为两类。第一类是非估算性的，这方面迄今最好的结果是1966年由中国数学家陈景润利用筛法(sieve method)所取得的②。陈景润证明了：存在无穷多个素数 $p$，使得 $p+2$ 要么是素数，要么是两个素数的乘积。这个结果的形式与他关于哥德巴赫猜想的结果很类似③。目前一般认为，由于筛法本身所具有的局限性，这一结果在筛法的范围之内已很难被超越。

证明孪生素数猜想的另一类结果则是估算性的，戈德斯通和伊尔迪里姆所取得的结果就属于这一类。这类结果估算的是相邻素数之间的最小间隔，更确切地说是：

$$\Delta = \liminf_{n\to\infty} \frac{p_{n+1} - p_n}{\ln p_n}$$

翻译成白话文，这个表达式所定义的是两个相邻素数之间的间隔与其中较小的那个素数的对数值之比在整个素数集合中所取的最小值。很明显，孪生素数猜想要想成立，$\Delta$ 必须为0。因为孪生素数猜想表明 $p_{n+1}-p_n=2$ 对无穷多个 $n$ 成立，而 $\ln(p_n)\to\infty$，因此两者之比的最小值对于孪生素数集合——从而对于整个素数集合也——趋于零。不过要注意，$\Delta=0$ 只是孪生素数猜想成立的**必要条件**，而不是充分条件。换句话说，如果能证明 $\Delta\neq0$，则孪生素数猜想就被推翻了；但证明了 $\Delta=0$，却并不意味着孪生素数猜想一定成立。

---

① 截至2011年底，这一纪录已被刷新为了：$(3\,756\,801\,695\,685\times2^{666\,669}-1,\ 3\,756\,801\,695\,685\times2^{666\,669}+1)$，这对素数中的每一个都长达200 700位。

② 顺便说一下，美国数学研究所在介绍本文开头所提到的戈德斯通和伊尔迪里姆的结果的简报中提到陈景润时所用的称呼是“伟大的中国数学家陈”(the great Chinese mathematician Chen)。

③ 陈景润关于哥德巴赫猜想的结果——被称为陈氏定理(Chen's theorem)——是：任何足够大的偶数都可以表示成两个数的和，其中一个是素数，另一个要么是素数，要么是两个素数的乘积。

对 Δ 最简单的估算来自于素数定理。按照素数定理，对于足够大的 $x$，在 $x$ 附近素数出现的几率为 $1/\ln(x)$，这表明素数之间的平均间隔为 $\ln(x)$，从而 $(p_{n+1}-p_n)/\ln(p_n)$ 给出的其实是相邻素数之间的间隔与平均间隔的比值，其平均值显然为 1①。平均值为 1，最小值显然是小于等于 1，因此素数定理给出 $\Delta\leqslant 1$。

对 Δ 的进一步估算始于哈代和李特伍德。1926 年，他们运用圆法（circle method）证明了假如广义黎曼猜想（generalized Riemann hypothesis）成立，则 $\Delta\leqslant 2/3$。这一结果后来被苏格兰数学家兰金（Robert Alexander Rankin）改进为 $\Delta\leqslant 3/5$。但这两个结果都有赖于本身尚未得到证明的广义黎曼猜想，因此只能算是有条件的结果。1940 年，匈牙利数学家埃尔德什（Paul Erdös）利用筛法率先给出了一个不带条件的结果：$\Delta<1$（即把素数定理给出的结果中的等号部分去掉了）。此后意大利数学家里奇（Giovanni Ricci）于 1954 年，意大利数学家蓬皮埃利（Enrico Bombieri）、英国数学家达文波特（Harold Davenport）于 1966 年，以及英国数学家赫克斯利（Martin Huxley）于 1977 年，分别将 Δ 的估算值推进到了 $\Delta\leqslant 15/16$，$\Delta\leqslant(2+\sqrt{3})/8\approx 0.4665$，以及 $\Delta\leqslant 0.4425$。戈德斯通和伊尔迪里姆之前最好的结果则是德国数学家梅尔（Helmut Maier）于 1986 年得到的 $\Delta\leqslant 0.2486$。

以上这些结果都是在小数点后面做文章，戈德斯通和伊尔迪里姆的结果将这一系列努力大大推进了一步，并且——如果得到确认的话——将在一定意义上终结对 Δ 进行数值估算的长达几十年的漫漫征途。因为戈德斯通和伊尔迪里姆所证明的结果是 $\Delta=0$。当然，如我们前面所述，$\Delta=0$ 只是孪生素数猜想成立的必要条件，而不是充分条件，因此戈德斯通和伊尔迪里姆的结果即便得到确认，离最终证明孪生素数猜想仍有相当的距离，但它无疑将是近十几年来这一领域中最引人注目的结果。

一旦 $\Delta=0$ 被证明，下一个努力方向会是什么呢？一个很自然的方向将是研究 Δ 趋于 0 的方式。孪生素数猜想要求 $\Delta\sim[\ln(p_n)]^{-1}$（因为 $p_{n+1}-$

① 这个“归一”性也正是在 Δ 的表达式中引进 $\ln(p_n)$ 的原因。

$p_n=2$ 对无穷多个 $n$ 成立)。戈德斯通和伊尔迪里姆的结果所给出的则是 $\Delta\sim[\ln(p_n)]^{-1/9}$,两者之间还有不小的差距①。但是看过戈德斯通和伊尔迪里姆手稿的一些数学家认为,戈德斯通和伊尔迪里姆所用的方法还存在改进空间。也就是说,他们的方法还有可能对 $\Delta$ 趋于 0 的方式作出更强的估计。从这个意义上讲,戈德斯通和伊尔迪里姆这一结果的价值不仅仅在于结果本身,更在于它有可能成为一系列未来研究的起点。这种带传承性的系列研究对于数学来说有着双重的重要性,因为一方面,这种研究可能取得的新结果将是对数学的直接贡献;另一方面,这种研究对戈德斯通和伊尔迪里姆的结果会起到反复推敲与核实的作用。现代数学早已超越了一两个评审花一两个小时就可以对一个数学证明做出评判的时代。著名的四色定理(four color theorem)和费马大定理(Fermat's Last Theorem)都曾有过一个证明时隔几年、甚至十几年才被发现错误的例子。因此,一个复杂的数学结果能成为进一步研究的起点,吸引其他数学家的参与,对于最终判定其正确性具有极其正面的意义②。

2003 年 4 月 6 日写于纽约

2014 年 9 月 15 日最新修订

---

① 本文发布之后,关于戈德斯通和伊尔迪里姆的工作又有了一些重要的后续发展,其中包括:2003 年 4 月 23 日,英国数学家格兰维尔(Andrew Granville)和印度数学家桑德拉拉扬(Kannan Soundararajan)发现了戈德斯通和伊尔迪里姆原始证明中的一个错误,并得到了戈德斯通和伊尔迪里姆的承认;2005 年初,戈德斯通和伊尔迪里姆"伙同"匈牙利数学家平兹(János Pintz)"卷土重来",再次证明了 $\Delta=0$。他们所证明的 $\Delta$ 的新的渐近行为是:$\Delta\sim[\ln\ln(p_n)]^2/[\ln(p_n)]^{1/2}$。

② 2013 年 5 月 14 日,《自然》(*Nature*)等科学杂志及大量中外媒体报道了旅美数学家张益唐在孪生素数猜想研究中所取得的一个重要的新进展,即证明了存在无穷多个素数对,其间隔小于 7 000 万。这一进展——如果得到确认的话——相当于证明了波利尼亚克猜想至少对某个小于 3 500 万的 $k$ 成立。用 $\Delta$ 来表述的话,则相当于不仅证明了 $\Delta=0$,而且给出了与孪生素数猜想所要求的相同的渐近行为:$\Delta\sim[\ln(p_n)]^{-1}$(不过,这一渐近行为跟 $\Delta=0$ 一样,只是孪生素数猜想成立的必要条件,而不是充分条件)。张益唐的证明用到了戈德斯通、平兹、伊尔迪里姆等人的结果,并于 2013 年 5 月 21 日被《数学年刊》(*Annals of Mathematics*)所接受。张益唐的结果也存在改进空间,截至 2014 年 3 月,陶哲轩等数学家已将其中的 7 000 万这一素数间隔"压缩"到了 246。

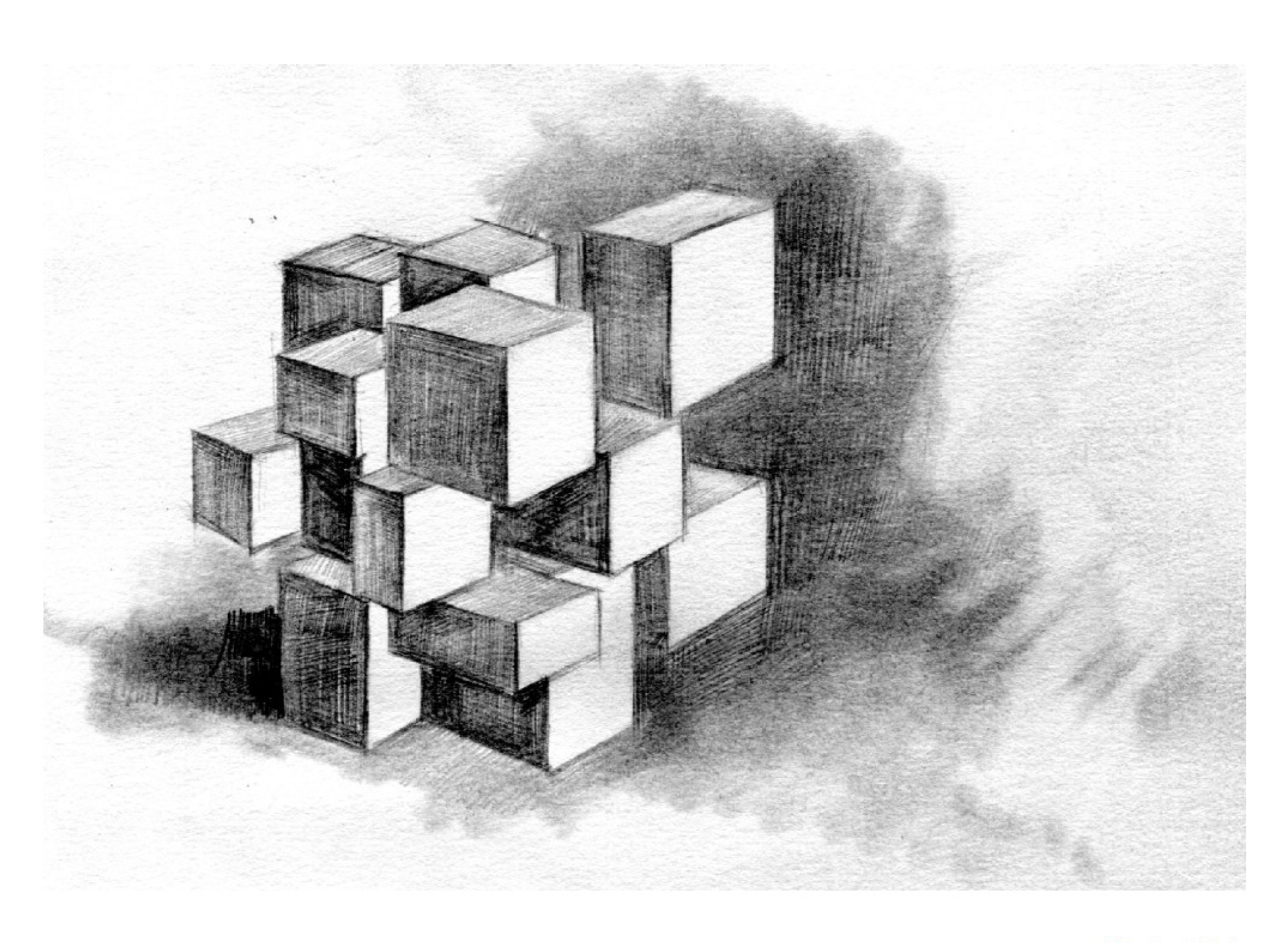
# 魔方与“上帝之数”[①]

2008 年 7 月，来自世界各地的很多优秀的魔方玩家聚集在捷克共和国(Czech Republic)中部城市帕尔杜比采(Pardubice)，参加魔方界的重要赛事：捷克公开赛。在这次比赛上，荷兰玩家阿克斯迪杰克(Erik Akkersdijk)创下了一个惊人的纪录：只用 7.08 秒就复原了一个颜色被打乱的魔方。无独有偶，在这一年的 8 月，人们在研究魔方背后的数学问题上也取得了重要进展。在本文中，我们就来介绍一下魔方以及它背后的数学问题。

## 一、风靡世界的玩具

1974 年春天，匈牙利布达佩斯应用艺术学院(Budapest College of Applied Arts)的建筑学教授鲁比克(Ernö Rubik)萌生了一个有趣的念头，那就是设计一个教学工具来帮助学生直观地理解空间几何中的各种转动。经过思考，他决定制作一个由一些小方块组成的，各个面能随意转动的 3×3×3 的立方体。这样的立方体可以很方便地演示各种空间转动。

① 本文曾发表于《科学画报》2008 年第 12 期(上海科学技术出版社出版)。



绘画：张京

这个想法虽好，实践起来却面临一个棘手的问题，即如何才能让这样一个立方体的各个面能随意转动？鲁比克想了很多点子，比如用磁铁或橡皮筋连接各个小方块，但都不成功。那年夏天的一个午后，他在多瑙河畔乘凉，眼光不经意地落在了河畔的鹅卵石上。忽然，他心中闪过一个新的设想：用类似于鹅卵石表面那样的圆形表面来处理立方体的内部结构。这一新设想成功了，鲁比克很快完成了自己的设计，并向匈牙利专利局申请了专利。这一设计就是我们都很熟悉的魔方（magic cube），也叫鲁比克方块（Rubik's cube）①。

6 年后，鲁比克的魔方经过一位匈牙利商人兼业余数学家的牵头，打进了西欧及美国市场，并以惊人的速度成为了风靡全球的新潮玩具。在此后的 25 年间，魔方的销量超过了 3 亿个。在魔方的玩家中，既有牙牙学语的孩子，也有跨国公司的老总。魔方虽未如鲁比克设想的那样成为一种空间几何的教学工具，却变成了有史以来最畅销的玩具。

魔方之畅销，最大的魔力就在于其数目惊人的颜色组合。一个魔方出厂时每个面各有一种颜色，总共有 6 种颜色，但这些颜色被打乱后，所能形成的组合数却多达 4 325 亿亿②（1 亿亿 $=1\times10^{16}$）。如果我们将这些组合中的每

---

① “魔方”是鲁比克自己为这一设计所取的名字，“鲁比克方块”则是美国玩具公司 Ideal Toys 所取的名字。在西方国家，鲁比克方块这一名称更为流行，在中国，则是魔方这一名称更为流行。另外要提醒读者的是，魔方有很多种类，本文介绍的 $3\times3\times3$ 魔方只是其中最常见的一种。

② 具体的计算是这样的：在组成魔方的小立方体中，有 8 个是顶点，它们之间有 8！种置换；这些顶点每个有 3 种颜色，从而在朝向上有 $3^7$ 种组合（由于结构所限，魔方的顶点只有 7 个能有独立朝向）。类似地，魔方有 12 个小立方体是边，它们之间有 12！/2 种置换（之所以除以 2，是因为魔方的顶点一旦确定，边的置换就只有一半是可能的）；这些边每个有两种颜色，在朝向上有 $2^{11}$ 种组合（由于结构所限，魔方的边只有 11 个能有独立朝向）。因此，魔方的颜色组合总数为 $8!\times3^7\times12!\times2^{11}/2=43\ 252\ 003\ 274\ 489\ 856\ 000$，即大约 4 325 亿亿。另外值得一提的是，倘若我们允许将魔方拆掉重组，则前面提到的结构限定将不复存在，它的颜色组合数将多达 51 900 亿亿种。不过颜色组合数的增加并不意味着复原的难度变大，魔方结构对颜色组合数的限制实际上正是使魔方的复原变得困难的主要原因。举个例子来说，26 个英文字母在相邻字母的交换之下共有约 400 亿亿亿种组合，远远多于魔方颜色的组合数，但通过相邻字母的交换将随意排列的 26 个英文字母复原成从 A 到 Z 的初始排列却非常简单。

一种都做成一个魔方，这些魔方排在一起，可以从地球一直排到250光年外的遥远星空——也就是说，如果我们在这样一排魔方的一端点上一盏灯，那灯光要在250年后才能照到另一端！如果哪位勤勉的玩家想要尝试所有的组合，哪怕他不吃、不喝、不睡，每秒钟转出10种不同的组合，也要花1 500亿年的时间才能如愿(作为比较，我们的宇宙目前还不到140亿岁)。与这样的组合数相比，广告商们常用的“成千上万”、“数以亿计”、“数以十亿计”等平日里虚张声势、忽悠顾客的形容词反倒变成了难得的谦虚。我们可以很有把握地说，假如不掌握诀窍地随意乱转，一个人哪怕从宇宙大爆炸之初就开始玩魔方，也几乎没有任何希望将一个色彩被打乱的魔方复原。

## 二、魔方与“上帝之数”

魔方的玩家多了，相互间的比赛自然是少不了的。自1981年起，魔方爱好者们开始举办世界性的魔方大赛，从而开始缔造自己的世界纪录。这一纪录被不断地刷新着，截至2013年，复原魔方的最快纪录已经达到了令人吃惊的5.55秒。当然，单次复原的纪录存在一定的偶然性，为了减少这种偶然性，自2003年起，魔方大赛的冠军改由多次复原的平均成绩来决定①，截至2013年，这一平均成绩的世界纪录为6.54秒。这些纪录的出现，表明魔方虽有天文数字般的颜色组合，但只要掌握窍门，将任何一种给定的颜色组合复原所需的转动次数却很可能并不多。

那么，最少需要多少次转动，才能确保**无论什么样的颜色组合**都能被复原呢②？这个问题引起了很多人，尤其是数学家们的兴趣。这个**复原任意组合**

① 确切地说是取5次尝试中居中的3次成绩的平均值。

② 为了使这一问题有意义，当然首先要定义什么是转动。在对魔方的数学研究中，转动是指将魔方的任意一个(包含9个小方块的)面沿顺时针或逆时针方向转动90°或180°，对每个面来说，这样的转动共有3种。(请读者想一想，为什么不是4种？)由于魔方有6个面，因此它的基本转动方式共有18种。

所需的最少转动次数被数学家们戏称为“上帝之数”(God's number)，而魔方这个玩具世界的宠儿则由于这个“上帝之数”而一举侵入了学术界。

要研究“上帝之数”，首先当然要研究魔方的复原方法。在玩魔方的过程中，人们早就知道，将任何一种**给定的颜色组合**复原都是很容易的，这一点已由玩家们的无数杰出纪录所示范。不过魔方玩家们所用的复原方法是便于人脑掌握的方法，却不是转动次数最少的，因此无助于寻找“上帝之数”。寻找转动次数最少的方法是一个有一定难度的数学问题。当然，这个问题是难不倒数学家的。早在 20 世纪 90 年代中期，人们就有了较实用的算法，可以用平均 15 分钟左右的时间找出复原一种**给定的颜色组合**的**最少转动次数**。从理论上讲，如果有人能对每一种颜色组合都找出这样的最少转动次数，那么这些转动次数中最大的一个无疑就是“上帝之数”了。但可惜的是，“4 325 亿亿”这个巨大数字成为了人们窥视“上帝之数”的拦路虎。如果采用上面提到的算法，用上面提到的速度寻找，哪怕用 1 亿台计算机同时进行，也要用超过 1 000 万年的时间才能完成。

看来蛮干是行不通的，数学家们于是便求助于他们的老本行：数学。从数学的角度看，魔方的颜色组合虽然千变万化，其实都是由一系列基本操作——即转动——产生的，而且那些操作还具有几个非常简单的特点，比如任何一个操作都有一个相反的操作(比如与顺时针转动相反的操作就是逆时针转动)。对于这样的操作，数学家们的“武器库”中有一种非常有效的工具来对付它，这工具叫做群论(group theory)，它比魔方早 140 多年就已出现了。据说德国数学大师希尔伯特(David Hilbert)曾经表示，学习群论的窍门就是选取一个好的例子。自魔方问世以来，数学家们已经写出了好几本通过魔方讲述群论的书。因此，魔方虽未成为空间几何的教学工具，却在一定程度上可以作为学习群论的“好的例子”。

对魔方研究来说，群论有一个非常重要的优点，就是可以充分利用魔方的对称性。我们前面提到“4 325 亿亿”这个巨大数字时，其实有一个疏漏，那就是未曾考虑到魔方作为一个立方体所具有的对称性。由此导致的结果，是那

4 325亿亿种颜色组合中有很多其实是完全相同的，只是从不同的角度去看——比如让不同的面朝上或者通过镜子去看——而已。因此，“4 325亿亿”这个令人望而生畏的数字实际上是“注水猪肉”。那么，这“猪肉”中的“水分”占多大比例呢？说出来吓大家一跳：占了将近99%！换句话说，仅凭对称性一项，数学家们就可以把魔方的颜色组合减少两个数量级①。

但减少两个数量级对于寻找“上帝之数”来说还是远远不够的，因为那不过是将前面提到的1000万年的时间减少为了10万年。对于解决一个数学问题来说，10万年显然还是太长了，而且我们也并不指望真有人能动用1亿台计算机来计算“上帝之数”。数学家们虽然富有智慧，在其他方面却不见得富有，他们真正能动用的也许只有自己书桌上那台计算机。因此为了寻找“上帝之数”，人们还需要更巧妙的思路。幸运的是，群论这一工具的威力远不只是用来分析像立方体的对称性那样显而易见的东西，在它的帮助下，更巧妙的思路很快就出现了。

## 三、寻找“上帝之数”

1992年，德国数学家科先巴(Herbert Kociemba)提出了一种寻找魔方复原方法的新思路②。他发现，在魔方的基本转动方式中，有一部分可以自成系列，通过这部分转动可以形成将近200亿种颜色组合③。利用这200亿种颜

---

① 确切地说，是减少至1/96，或45亿亿种组合。

② 科先巴的新思路是本文介绍的一系列计算研究的起点，但并不是最早的魔方算法。早在1981年，目前在美国田纳西大学(University of Tennessee)，当时在伦敦南岸大学(London South Bank University)的数学家西斯尔斯韦特(Morwen Thistlethwaite)就提出了一种算法，被称为西斯尔斯韦特算法(Thistlethwaite algorithm)。西斯尔斯韦特算法可保证通过不超过52次转动复员魔方的任意一种颜色组合(相当于证明了上帝之数不超过52)，在科先巴新思路问世之前的1991年，这一数字曾被压缩到42。

③ 确切地说，是18种基本转动方式中有10种自成系列，由此形成的颜色组合共有8！×8！×4！/2(约195亿)种。

色组合，科先巴将魔方的复原问题分解成了两个步骤：第一步是将任意一种颜色组合转变为那200亿种颜色组合之一，第二步则是将那200亿种颜色组合复原。如果我们把魔方的复原比作是让一条汪洋大海中的小船驶往一个固定目的地，那么科先巴提出的那200亿种颜色组合就好比是一片特殊水域——一片比那个固定目的地大了200亿倍的特殊水域。他提出的两个步骤就好比是让小船首先驶往那片特殊水域，然后从那里驶往那个固定目的地。在汪洋大海中寻找一片巨大的特殊水域，显然要比直接寻找那个小小的目的地容易得多，这就是科先巴新思路的巧妙之处。

但即便如此，要用科先巴的新思路对"上帝之数"进行估算仍不是一件容易的事。尤其是，要想进行快速计算，最好是将复原那200亿种颜色组合的最少转动次数（这相当于是那片特殊水域的"地图"）存储在计算机的内存中，这大约需要300兆（300MB）的内存。300兆在今天看来是一个不太大的数目，但在科先巴提出新思路的年代，普通计算机的内存连它的十分之一都远远不到。因此直到3年之后，才有人利用科先巴的新思路给出了第一个估算结果。此人名叫里德（Michael Reid），是美国中佛罗里达大学（Unversity of Central Florida）的数学家。1995年，里德通过计算发现，最多经过12次转动，就可以将魔方的任意一种颜色组合转变为科先巴新思路中那200亿种颜色组合之一；而最多经过18次转动，就可以将那200亿种颜色组合中的任意一种复原。这表明，最多经过12＋18＝30次转动，就可以将魔方的任意一种颜色组合复原。

在得到上述结果后，里德很快对自己的估算作了改进，将结果从30减少为了29，这表明"上帝之数"不会超过29。此后随着计算机技术的发展，数学家们对里德的结果又作出了进一步改进，但进展并不迅速。直到11年后的2006年，奥地利开普勒大学（Johannes Kepler University）符号计算研究所（Research Institute for Symbolic Computation）的博士生拉杜（Silviu Radu）才将结果推进到了27。第二年（即2007年），美国东北大学（Northeastern University）的计算机科学家孔克拉（Dan Kunkle）和库伯曼（Gene

Cooperman)又将结果推进到了 26,他们的工作采用了并行计算系统,所用的最大存储空间高达 700 万兆($7\times10^6$ MB),所耗的计算时间则长达 8 000 小时(相当于将近一年的 24 小时不停歇计算)。

这些计算表明,“上帝之数”不会超过 26。但是,所有这些计算的最大优点——即利用科先巴新思路中那片特殊水域——同时也是它们最致命的弱点,因为它们给出的复原方法都必须经过那片特殊水域。可事实上,很多颜色组合的最佳复原方法根本就不经过那片特殊水域,比如紧邻目的地,却恰好不在特殊水域中的任何小船,显然都没必要像中国大陆和台湾之间的直航包机一样,故意从那片特殊水域绕一下才前往目的地。因此,用科先巴新思路得到的复原方法未必是最佳的,由此对“上帝之数”所做的估计也极有可能是高估。

可是,如果不引进科先巴新思路中的特殊水域,计算量又实在太大,怎么办呢?数学家们决定采取折中手段,即扩大那片特殊水域的“面积”。因为特殊水域越大,最佳复原路径恰好经过它的可能性也就越大(当然,计算量也会有相应的增加)。2008 年,研究“上帝之数”长达 15 年之久的计算机高手罗基奇(Tomas Rokicki)运用了相当于将科先巴新思路中的特殊水域扩大几千倍的巧妙方法,在短短几个月的时间内对“上帝之数”连续发动了四次猛烈攻击,将它的估计值从 25 一直压缩到了 22——这就是本文开头提到的人们在研究魔方背后的数学问题上取得的重要进展。罗基奇的计算得到了电影特效制作商索尼图形图像运作公司(Sony Pictures Imageworks)的支持,这家曾为《蜘蛛人》(*Spider-Man*)等著名影片制作特效的公司向罗基奇提供了相当于 50 年不停歇计算所需的计算机资源。

由此我们进一步知道,“上帝之数”一定不会超过 22。但是,罗基奇虽然将科先巴新思路中的特殊水域扩展得很大,终究仍有一些颜色组合的最佳复原方法是无需经过那片特殊水域的,因此,“上帝之数”很可能比 22 更小。那么,它究竟是多少呢?种种迹象表明,它极有可能是 20。这是因为,人们在过去这么多年的所有努力——其中包括罗基奇直接计算过的大约 4 000 万亿种

颜色组合——中，都从未遇到过任何必须用20次以上转动才能复原的颜色组合，这表明“上帝之数”很可能不大于20；另一方面，人们已经发现了几万种颜色组合，它们必须要用20次转动才能复原，这表明“上帝之数”不可能小于20。将这两方面综合起来，数学家们普遍相信，“上帝之数”的真正数值就是20。

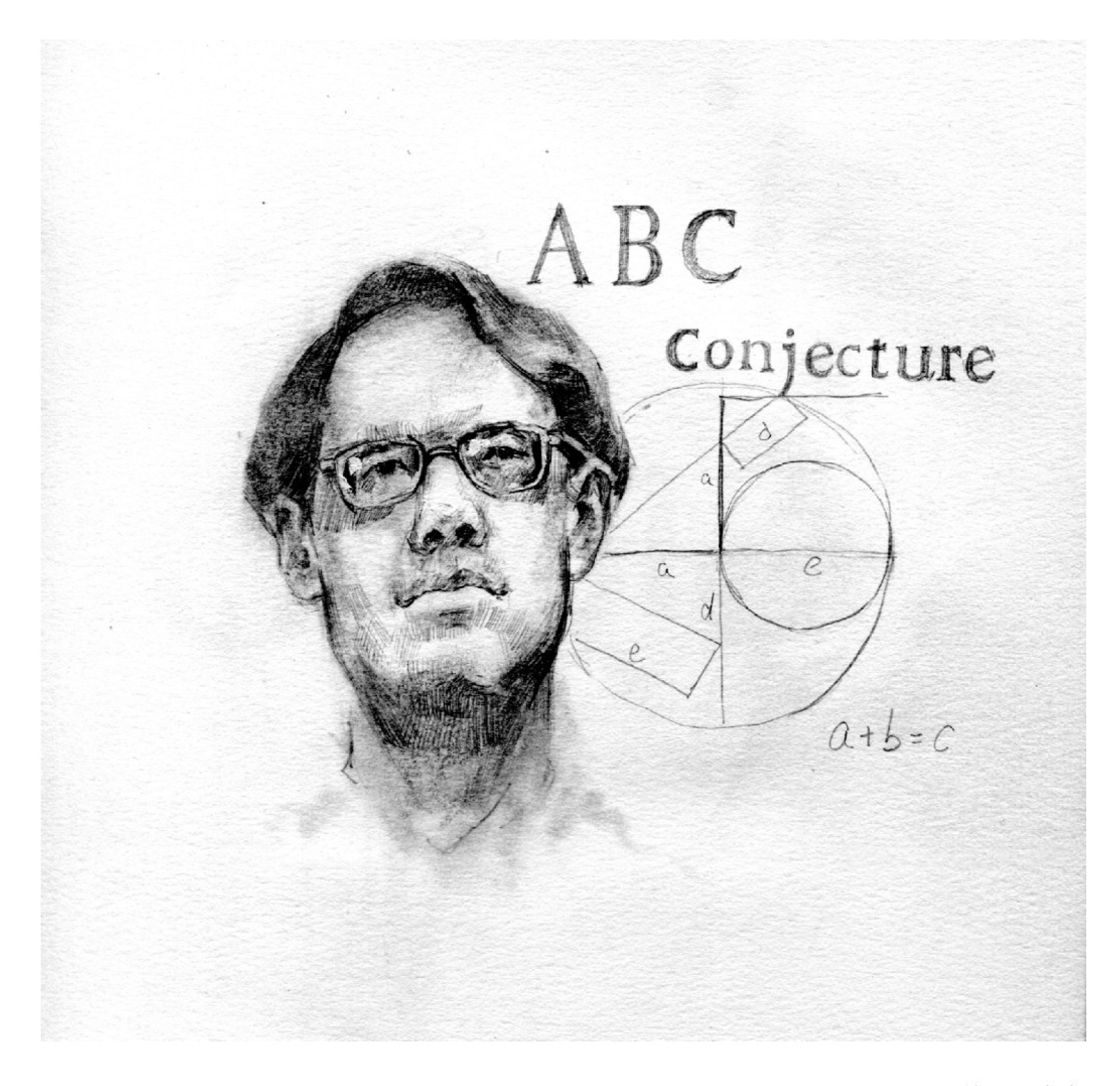
2010年8月，这个游戏与数学交织而成的神秘的“上帝之数”终于水落石出：研究“上帝之数”的“元老”科先巴、“新秀”罗基奇，以及另两位合作者——戴维森（Morley Davidson）和德斯里奇（John Dethridge）——宣布了对“上帝之数”是20的证明[①]。他们的证明得到了谷歌公司（Google）提供的相当于英特尔（Intel）四核心处理器35年不停歇计算所需的计算机资源。

因此，现在我们可以用数学特有的确定性来宣布“上帝之数”的数值了，那就是：20。

2008年11月2日写于纽约

2014年9月18日最新修订

① 他们所宣布的证明完成时间为2010年7月。




绘画：张京

# ABC 猜想浅说①

由前三个英文字母拼合而成的“ABC”一词据说自 13 世纪起便见诸文献了，含义为“入门”。这些年随着英文在中国的流行，该词在中文世界里也夺得了一席之地，出现在了很多图书的书名中，大有跟中文词“入门”一较高下之势。不过，倘若你在数学文献中看到一个以“ABC”命名的猜想——“ABC 猜想”（ABC conjecture），千万不要以为那是一个“入门”级别的猜想。事实上，这一猜想在公众知名度方面或许尚处于“入门”阶段，以难度和地位而论却绝不是“入门”级别的。

在本文中，我们将对这一并非“入门”级别的猜想做一个“入门”级别的介绍。

## 一、什么是 ABC 猜想？

在介绍之前，让我们先回忆一下中小学数学中的两个简单概念。其中第一个概念是素数（prime number）。我们知道，很多正整数可以分解为其他——

① 本文是应《南方周末》约稿而写的“ABC 猜想”简介，曾以《望月“摘月”》为标题发表于 2012 年 10 月 25 日（发表稿经编辑改动，系删节版）。本文的完整版发表于《数学文化》2014 年 11 月刊。

即不同于它自己的——正整数的乘积,比如 9=3×3,231=3×7×11,等等。但也有一些正整数不能这么分解,比如 13,29 等。这后一类正整数——1 除外——就是所谓的素数。素数是一个被称为“数论”(number theory)的数学分支中的核心概念,其地位常被比喻为物理学中的原子(atom),因为与物理学中物质可以分解为原子相类似,数学中所有大于 1 的正整数都可以分解为素数的乘积(素数本身被视为是自己的分解)①。第二个概念则是互素(co-prime)。两个正整数如果其素数分解中不存在共同的素数,就称为是互素的,比如 21=3×7 和 55=5×11 就是互素的②。

有了这两个简单概念,我们就可以介绍 ABC 猜想了。ABC 猜想针对的是满足两个简单条件的正整数组($A$,$B$,$C$)③。其中第一个条件是 $A$ 和 $B$ 互素,第二个条件是 $A+B=C$。显然,满足这种条件的正整数组——比如(3, 8, 11)、(16, 17, 33)……——有无穷多个(请读者自行证明)。为了引出 ABC 猜想,让我们以(3, 8, 11)为例,做一个“三步走”的简单计算:

(1) 将 $A$、$B$、C 乘起来(结果是 3×8×11=264);

(2) 对乘积进行素数分解(结果是 $264=2^3\times3\times11$);

(3) 将素数分解中所有不同的素数乘起来(结果是 2×3×11=66)。

现在,让我们将 $A$、$B$、$C$ 三个数字中较大的那个(即 $C$)与步骤(3)的结果比较一下。我们发现后者大于前者(因为后者为 66,前者为 11)。读者可以对上面所举的另一个例子——即(16, 17, 33)——也试一下,你会发现同样的结果。如果随便找一些其他例子,你也很可能发现同样的结果。

---

① 不仅如此,这样的分解还可以被证明是唯一的,这被称为算术基本定理(fundamental theorem of arithmetic)。

② 对这一定义还有一个小小的补充,即 1 被定义为与所有正整数都互素。

③ 为了简单起见,我们的介绍是针对正整数的,但 ABC 猜想其实也可以针对整数进行表述,两者并无实质差别。我们将后者留给感兴趣的读者去做。

但你若因此以为这是规律，那就完全错了，因为它不仅不是规律，而且有无穷多的反例。比如(3，125，128)就是一个反例(请读者自行验证)。但是，数学家们猜测，如果把步骤(3)的结果放大成它的一个大于1的幂，那个幂哪怕只比1大上一丁点儿(比如1.000 000 000 01)，情况就有可能大不一样。这时它虽仍未必保证能够大于三个数字中较大的那个(即$C$)，但反例的数目将由无穷变为有限。这个猜测就是所谓的ABC猜想①，它是由英国数学家麦瑟尔(David Masser)和法国数学家厄斯特勒(Joseph Oesterlé)于20世纪80年代中期彼此独立地提出的。“ABC”这个毫无创意的名字——大家可能猜到了——则是来自把猜想中涉及到的三个数字称为$A$、$B$、$C$的做法，而非“入门”之意。

与数学猜想大家庭中的著名成员，如黎曼猜想(Riemann hypothesis)、哥德巴赫猜想(Goldbach conjecture)、孪生素数猜想(twin prime conjecture)，以及(已被证明了的)曾经的费马猜想(Fermat conjecture)、四色猜想(four-color conjecture)等相比，ABC猜想的“资历”是很浅的(其他那些猜想都是百岁以上的“老前辈”)，公众知名度也颇有不及，但以重要性而论，则除黎曼猜想外，上述其他几个猜想都得退居其后。

---

① 这里可以略作一点补充：步骤(3)的结果因不含任何素数因子的平方，被称为$A$、$B$、$C$三个数字乘积的“无平方部分”(square-free part)，简记为sqp($ABC$)——不过要注意的是，这一记号在某些文献中有不同含义，与本文含义相一致的另一种记号为rad($ABC$)。用这一记号，ABC猜想可以表述为“对任意给定的$n>1$，只有有限多组($A$,$B$,$C$)满足$\mathrm{sqp}(ABC)^n<C$”(当然，别忘了$A$和$B$互素及$A+B=C$这两个条件)。这一表述通常见诸科普介绍，在专业文献中ABC猜想往往被表述为“对任意给定的$n>1$，$\mathrm{sqp}(ABC)^n/C$的下界大于零”。感兴趣的读者不妨由“科普表述”出发，证明一下“专业表述”(不过要提醒读者的是：相反方向的证明，即由“专业表述”证明“科普表述”，并不是轻而易举的)。另外要说明的是，正文提到的所谓ABC猜想所允许的“反例”乃是“科普表述”特有的提法，意指满足$\mathrm{sqp}(ABC)^n<C$的那有限多组($A$,$B$,$C$)，在“专业表述”中是没有所谓“反例”的提法的。

## 二、ABC 猜想为什么重要?

ABC 猜想有一个在普通人看来并不奥妙的特点,就是将整数的加法性质(比如 $A+B=C$)和乘法性质(比如素数概念——因为它是由乘法性质所定义的)交互在了一起。不过,数学家们早就知道,由这两种本身很简单的性质交互所能产生的复杂性是近乎无穷的。数论中许多表述极为浅显,却极难证明的猜想(或曾经的猜想),比如前面提到的哥德巴赫猜想、孪生素数猜想、费马猜想等都具有这种加法性质和乘法性质相交互的特性。数论中一个很重要的分支——旨在研究整系数代数方程的整数解的所谓丢番图分析(Diophantine analysis)——更是整个分支都具有这一特性。丢番图分析的困难性是颇为出名的,著名德国数学家希尔伯特(David Hilbert)曾乐观地希望能找到其"一揽子"的解决方案,可惜这个被称为希尔伯特第十问题的希望后来落了空,被证明是不可能实现的。[①] 与希尔伯特的乐观相反,美国哥伦比亚大学(Columbia University)的数学家戈德菲尔德(Dorian Goldfeld)曾将丢番图分析比喻为飞蝇钓(fly-fishing)——那是发源于英国贵族的一种特殊的钓鱼手法,用甩出去的诱饵模拟飞蝇等昆虫的飞行姿态,以吸引凶猛的掠食性鱼类。飞蝇钓的特点是技巧高、难度大、成功率低,而且只能一条一条慢慢地钓——象征着丢番图分析只能一个问题一个问题慢慢地啃,而无法像希尔伯特所希望的那样"一揽子"地解决掉。

但是,与交互了加法性质和乘法性质的其他猜想或问题不同的是,ABC 猜想这个从表述上看颇有些拖泥带水(因为允许反例)的猜想似乎处于某种中枢地位上,它的解决将直接导致一大类其他猜想或问题的解决。拿丢番图分析来说,戈德菲尔德就表示,假如 ABC 猜想能被证明,丢番图分析将由飞蝇钓变为最强力——乃至野蛮——的炸药捕鱼,一炸就是一大片,因为 ABC 猜想

① 对这一点感兴趣的读者可参阅拙作《小楼与大师:科学殿堂的人和事》(清华大学出版社,2014 年)中的《希尔伯特第十问题漫谈》一文。

能“将无穷多个丢番图方程转变为单一数学命题”。这其中最引人注目的“战利品”将是曾作为猜想存在了 300 多年，一度被《吉尼斯世界纪录》(*Guinness Book of World Records*)称为“最困难数学问题”的费马猜想。这个直到 1995 年才被英国数学家怀尔斯(Andrew Wiles)以超过 100 页的长篇论文所解决的猜想在 ABC 猜想成立的前提下，将只需不到一页的数学推理就能确立①。其他很多长期悬而未决的数学猜想或问题也将被“一锅端”。这种与其他数学命题之间的紧密联系是衡量一个数学命题重要性的首要“考评”指标，ABC 猜想在这方面无疑能得高分——或者用戈德菲尔德的话说，是“丢番图分析中最重要的未解决问题”，“是一种美丽”。

ABC 猜想的重要性吸引了很多数学家的兴趣，但它的艰深迟滞了取得进展的步伐。截至 2001 年，数学家们在这一猜想上取得的最好结果乃是将上述步骤(3)的结果放大成它的某种指数函数②。由于指数函数的大范围增长速度远比幂函数快得多，由它来保证其大于 $A$、$B$、$C$ 三个数字中较大的那个(即 $C$)当然要容易得多(相应地，命题本身则要弱得多)。

---

① 这个关于在 ABC 猜想成立的前提下，费马猜想将只需“不到一页的数学推理就能确立”(establishing in less than a page of mathematical reasoning)的不无夸张的说法出自美国数学协会(Mathematical Association of America)的出版主管、著名美国数学科普作家彼得森(Ivars Peterson)。不过，该说法虽然夸张，却并非完全“忽悠”。为了说明这一点，并作为对如何由 ABC 猜想证明其他命题的演示，我们在这里介绍一个“不到一页的数学推理”：假设费马猜想不成立，即存在互素的(这点请读者自行证明)正整数 $x$、$y$、$z$ 使得 $x^k+y^k=z^k$ ($k>2$)。则由前一条注释给出的 ABC 猜想的“专业表述”可知(取 $n=7/6$)：$\mathrm{sqp}(x^k y^k z^k)^{7/6}/z^k>\varepsilon(\varepsilon>0)$。由于 $\mathrm{sqp}(x^k y^k z^k)=\mathrm{sqp}(xyz)\leqslant xyz<z^3$，因此 $z^{3.5-k}>\varepsilon$。显然，对所有 $k\geqslant4$，只有小于(由 $\varepsilon$ 决定的)某个数值的有限多个 $z$ 能满足该不等式，而且当 $k$ 大于(由 $\varepsilon$ 决定的)某个数值后，将不会有任何 $z$ 满足该不等式。这表明，对所有 $k\geqslant4$，费马猜想的反例即便有也只能有有限多个，而且 $k$ 大到一定程度后将不再有反例。因此，证明费马猜想就变成了证明 $k=3$ 的情形(这在两百多年前就已完成)，以及通过数值验证排除总数有限的反例。这虽然并非“不到一页的数学推理”就能确立的，比起怀尔斯的证明来毕竟是直截了当多了。倘若历史走的是不同的路径，费马是在 ABC 猜想被证明之后才提出的费马猜想，他那句戏剧性的“我发现了一个真正出色的证明，可惜页边太窄写不下来”倒是不无成立之可能。

② 具体地说，截至 2001 年，这方面的最好结果是 $\exp[K\cdot\mathrm{sqp}(ABC)^{1/3+\varepsilon}]/C>1$，其中 $K$ 是与 $\varepsilon$ 有关(但与 $A$、$B$、$C$ 无关)的常数。

除上述理论结果外，自2006年起，由荷兰莱顿大学(Leiden University)的数学系牵头，一些数学和计算机爱好者建立了一个名为ABC@Home的分布式计算(distributed computing)系统，用以寻找ABC猜想所允许的反例。截至2014年4月，该系统已经找到了超过2 380万个反例，而且还在继续增加着。不过，与这一系统的著名“同行”——比如寻找外星智慧生物的SETI以及计算黎曼ζ函数非平凡零点的已经关闭了的ZetaGrid——不同的是，ABC@Home是既不可能证明，也不可能否证ABC猜想的(因为ABC猜想本就允许数量有限的反例)。从这个意义上讲，ABC@Home的建立更多地只是出于对具体反例——尤其是某些极端情形下的反例，比如数值最大的反例——的好奇。当然，具体反例积累多了，是否会衍生出有关反例分布的猜想，也是不无趣味的悬念。另外，ABC猜想还有一些拓展版本，比如对某些情形下的反例数目给出具体数值的版本，ABC@Home对那种版本原则上是有否证能力的。

## 三、ABC猜想被证明了吗?

如前所述，ABC猜想的公众知名度与一些著名猜想相比是颇有不及的。不过，2012年9月初，包括《自然》(*Nature*)、《科学》(*Science*)在内的一些重量级学术刊物，以及包括《纽约时报》(*New York Times*)在内的许多著名媒体却纷纷撰写或转载了有关ABC猜想的消息，使这一猜想在短时间内着实风光了一番。促成这一风光的是日本数学家望月新一(Shinichi Mochizuki)。2012年8月底，望月新一发表了由四篇长文组成的系列论文的第四篇，宣称证明了包括ABC猜想在内的若干重要猜想。这一宣称被一些媒体称为是能与1993年怀尔斯宣称证明了费马猜想，以及2002年佩雷尔曼(Grigory Perelman)宣称证明了庞加莱猜想(Poincaré conjecture)相提并论的事件。

由于这一原因，我应约撰写本文时，约稿编辑曾希望我能找认识望月新一的华人数学家聊聊，挖出点独家新闻来。可惜我不得不有负此托了，因为别说是我，就连《纽约时报》等擅挖材料的重量级媒体在报道望月新一其人时，也基

本没能超出他在自己网站上公布的信息。

按照那些信息，望月新一 1969 年 3 月 29 日出生于日本东京，16 岁（即 1985 年）进入美国普林斯顿大学（Princeton University）就读本科，三年后进入研究生院，师从著名德国数学家、1986 年菲尔茨奖（Fields Medal）得主法尔廷斯（Gerd Faltings），23 岁（即 1992 年）获得数学博士学位。此后，他先是“海归”成京都大学（Kyoto University）数理解析研究所（Research Institute for Mathematical Sciences）的研究助理（Research Associate），几个月后又前往美国哈佛大学从事了近两年的研究，然后重返京都大学。2002 年，33 岁的望月新一成为了京都大学数理解析研究所的教授。望月新一的学术声誉颇佳，曾获得过日本学术奖章（Japan Academy Medal）等荣誉。

有关望月新一其人的信息大体就是这些，但读者不必过于失望，因为望月新一所宣称的对 ABC 猜想的证明虽引起了很大关注，离公认还颇有距离，因此目前恐怕还未到挖掘其生平的最佳时机。事实上，在 ABC 猜想并不漫长的历史中，这并不是第一次有人宣称解决了这一猜想。2007 年，法国数学家施皮罗（Lucien Szpiro）就曾宣称解决了 ABC 猜想。施皮罗的学术声誉不在望月新一之下，不仅是领域内的专家，其工作甚至间接促成了 ABC 猜想的提出。但是，人们很快就在他的证明中发现了漏洞。这种宣称解决了一个重大数学猜想，随后却被发现漏洞的例子在数学史上比比皆是。因此，任何证明从宣称到公认，必须经过同行的严格检验。这一检验视证明的复杂程度而定，可长可短。不过对于望月新一的“粉丝”来说，恐怕得有长期等待的心理准备，因为望月新一那四篇论文的总长度超过了 500 页，几乎是怀尔斯证明费马猜想的论文长度的四倍！更糟糕的是，望月新一的证明采用了他自己发展起来的数学工具，这种工具据说是对以抽象和艰深著称的 1966 年菲尔兹奖得主格罗滕迪克（Alexander Grothendieck）的某些代数几何方法的推广，除他本人外，数学界并无第二人通晓①。就连研究方向与望月新一相近的英国牛津大学

① 望月新一自创的那种数学工具被称为 inter-universal Teichmuller theory 或 inter-universal geometry。他在其网站上则称自己为 Inter-universal Geometer。

(University of Oxford)的韩国数学家金明迥(Minhyong Kim)都表示,“我甚至无法对[望月新一的]证明给出一个专家概述,因为我并不理解它”,“仅仅对局势有一个一般了解也得花费一段时间”。美国威斯康星大学(University of Wisconsin)的数学家艾伦伯格(Jordan Ellenberg)则表示阅读望月新一的论文“仿佛是在阅读外星人的东西”(reading something from outer space)。2006 年菲尔茨奖得主、澳大利亚数学家陶哲轩(Terence Tao)也表示“现在对这一证明有可能正确还是错误做出评断还为时过早”。

像望月新一那样宣称用自创的数学工具证明著名数学猜想的事例在数学界也是有先例的。2004 年,美国普渡大学(Purdue)的数学教授德布朗基(Louisde Branges)宣称证明了著名的黎曼猜想,他所用的也是自创的数学工具。不过德布朗基在数学界的声誉和口碑均极差,加之年事已高(七旬老汉),其宣称遭到了数学界的冷淡对待①。与之不同的是,望月新一却不仅有良好的学术声誉,精力和研究能力也尚处于巅峰期。用陶哲轩的话说,望月新一“与佩雷尔曼和怀尔斯类似”,“是一个多年来致力于解决重要问题,在领域内享有很高声誉的第一流数学家”。有鉴于此,数学界不仅对望月新一的证明给予了重视,对他自创的方法也表示了兴趣,比如美国斯坦福大学(Stanford University)的数学家康拉德(Brian Conrad)就表示“激动人心之处不仅在于[ABC]猜想有可能已被解决,而且在于他[望月新一]必须引入的技巧和洞见应该是解决未来数论问题的非常有力的工具”。戈德菲尔德也认为“望月新一的证明如果成立,将是 21 世纪数学最惊人的成就”。

在这种兴趣的驱动下,一些数学家已经开始对望月新一的证明展开检验与讨论,比如著名数学讨论网站 Math Overflow 就已出现了一些有金明迥、陶哲轩等一流数学家参与的认真讨论。不过,检验过程何时才能完成,目前还不得而知,检验的结果如何,更是无从预料。证明得到公认固然是很多人乐意见到的,但一个长达 500 多页的证明存在漏洞也是完全可能的,当年怀尔斯对费

---

① 对此事感兴趣的读者可参阅拙作《黎曼猜想漫谈》的第 35 章。

马猜想的“只有”100 多页的证明，其早期版本就存在过漏洞，经过一年多的时间才得以弥补。不过，无论望月新一的证明是否成立，不少数学家对 ABC 猜想本身的成立倒是都抱有乐观态度，这一方面是因为能因这一猜想的成立而得到证明的很多数学命题(比如如今被称为费马大定理的费马猜想)已经通过其他途径得到了证明，从而表明 ABC 猜想的成立与数学的其他部分有很好的相容性(著名的黎曼猜想也有这样的特点)。另一方面，ABC 猜想还得到了一些启发性观点的支持，比如陶哲轩就从所谓的“概率启发式理由”(probabilistic heuristic justification)出发，预期 ABC 猜想应该成立①。

当然，信心和预期取代不了证明。望月新一证明的命运将会如何？ABC 猜想究竟被证明了没有？都将有待时间来回答②。

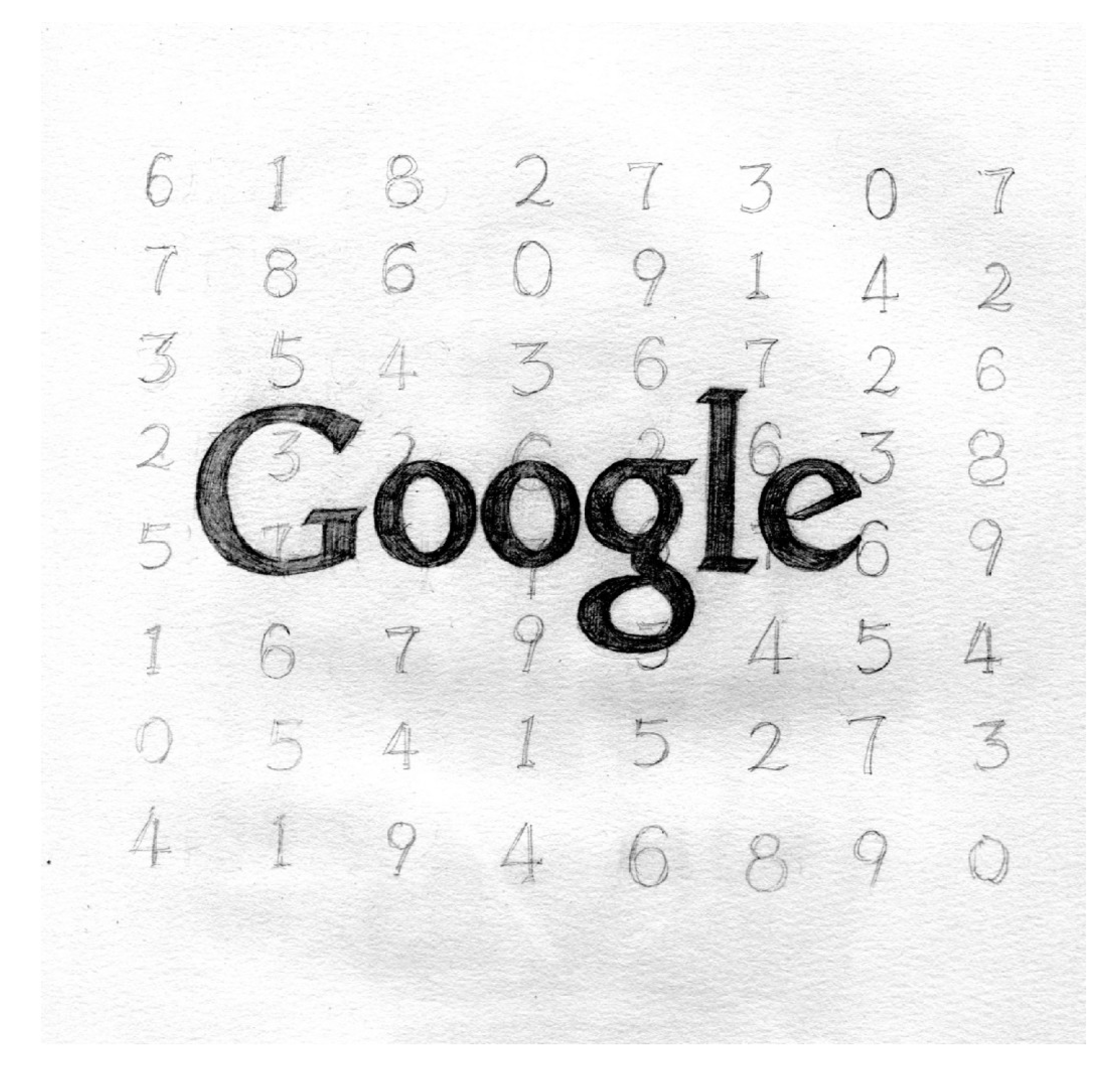
2012 年 10 月 14 日写于纽约

2014 年 10 月 1 日最新修订

---

① 陶哲轩的“概率启发式理由”的要点是将数论命题——比如一个数是素数——视为概率性命题，并利用概率工具来猜测数学命题的成立与否。这种做法的一个例子是对强孪生素数猜想成立的猜测(参阅收录于本书的拙作“孪生素数猜想”所介绍的有关该猜想的“简单的定性分析”)。

② 望月新一的证明发布至今已两年多，这期间美国耶鲁大学(Yale University)的数学系研究生季米特洛夫(Vesselin Dimitrov)及斯坦福大学(Stanford University)的数学家文卡塔斯(Akshay Venkatesh)曾写信向他指出过一个错误。望月新一承认了错误，但表示那是一个不影响结论的小错误。此后，他数度更新了自己的论文，截至本文修订之日(2014 年 10 月 1 日)，他更新后的四篇论文总长度超过了 550 页，最近一次更新的日期则为 2014 年 9 月 15 日。




绘画：张京

# 谷歌背后的数学[①]

## 一、引言

在如今这个互联网时代,有一家公司家喻户晓——它自 1998 年问世以来,在极短的时间内就声誉鹊起,不仅超越了所有竞争对手,而且彻底改观了整个互联网的生态。这家公司就是当今互联网上的第一搜索引擎:谷歌(Google)。

在这样一家显赫的公司背后,自然有许许多多商战故事,也有许许多多成功因素。但与普通商战故事不同的是,在谷歌的成功背后起着最关键作用的却是一个数学因素。

本文要谈的就是这个数学因素。

谷歌作为一个搜索引擎,它的核心功能顾名思义,就是网页搜索。说到搜索,我们都不陌生,因为那是凡地球人都会的技能。我们在字典里查个生字,在图书馆里找本图书,甚至在商店里寻一种商品,等等,都是搜索。只要稍稍推究一下,我们就会发现那些搜索之所以可能,并且人人都会,在很大程度上

① 本文曾发表于《数学文化》2011 年 2 月刊(山东大学与香港浸会大学合办)。

得益于以下三条：

（1）搜索对象的数量较小——比如一本字典收录的字通常只有一两万个，一家图书馆收录的不重复图书通常不超过几十万种，一家商店的商品通常不超过几万种，等等。

（2）搜索对象具有良好的分类或排序——比如字典里的字按拼音排序，图书馆里的图书按主题分类，商店里的商品按品种或用途分类，等等。

（3）搜索结果的重复度较低——比如字典里的同音字通常不超过几十个，图书馆里的同名图书和商店里的同种商品通常也不超过几十种，等等。

但互联网的鲜明特点却是以上三条无一满足。事实上，即便在谷歌问世之前，互联网上的网页总数就已超过了诸如图书馆藏书数量之类传统搜索对象的数目。而且这还只是冰山一角，因为与搜索图书时单纯的书名搜索不同，互联网上的搜索往往是对网页内容的直接搜索，这相当于将图书里的每一个字都变成了搜索对象，由此导致的数量才是真正惊人的，它不仅直接破坏了上述第一条，而且连带破坏了二、三两条。在互联网发展的早期，像雅虎（Yahoo）那样的门户网站曾试图为网页建立分类系统，但随着网页数量的激增，这种做法很快就“挂一漏万”了。而搜索结果的重复度更是以快得不能再快的速度走向失控。这其实是可以预料的，因为几乎所有网页都离不开几千个常用词，因此除非搜索生僻词，否则出现几十万、几百万、甚至几千万条搜索结果都是不足为奇的。

互联网的这些“不良特点”给搜索引擎的设计带来了极大的挑战。而在这些挑战之中，相对来说，对一、二两条的破坏是比较容易解决的，因为那主要是对搜索引擎的存储空间和计算能力提出了较高要求，只要有足够多的钱来买“装备”，这些都还能算是容易解决的——套用电视连续剧《蜗居》中某贪官的台词来说，“能用钱解决的问题就不是大问题”。但对第三条的破坏却要了命了，因为无论搜索引擎的硬件如何强大，速度如何快捷，要是搜索结果有几百万条，那么任何用户想从其中“海选”出自己真正想要的东西都是几乎不可能的。这一点对早期搜索引擎来说可谓是致命伤，而且它不是用钱就能解决的问题。

这致命伤该如何治疗呢？药方其实很简单，那就是对搜索结果进行排序，把用户最有可能需要的网页排在最前面，以确保用户能很方便地找到它们。但问题是：网页的水平千差万别，用户的喜好更是万别千差，互联网上有一句流行语叫做："在互联网上，没人知道你是一条狗（On the Internet，nobody knows you're a dog）。"连用户是人是狗都"没人知道"，搜索引擎又怎能知道哪些搜索结果是用户最有可能需要的，并对它们进行排序呢？

在谷歌主导互联网搜索之前，多数搜索引擎采用的排序方法，是以被搜索词语在网页中的出现次数来决定排序——出现次数越多的网页排在越前面。这个判据不能说毫无道理，因为用户搜索一个词语，通常表明对该词语感兴趣。既然如此，那该词语在网页中的出现次数越多，就越有可能表示该网页是用户所需要的。可惜的是，这个貌似合理的方法实际上却行不大通。因为按照这种方法，任何一个像祥林嫂一样翻来覆去倒腾某些关键词的网页，无论水平多烂，一旦被搜索到，都立刻会"金榜题名"，这简直就是广告及垃圾网页制造者的天堂。事实上，当时几乎没有一个搜索引擎不被"祥林嫂"们所困扰，其中最具讽刺意味的是：在谷歌诞生之前的 1997 年 11 月，堪称早期互联网巨子的当时四大搜索引擎在搜索自己公司的名字时，居然只有一个能使之出现在搜索结果的前十名内，其余全被"祥林嫂"们挤跑了。

## 二、基本思路

正是在这种情况下，1996 年初，谷歌公司的创始人，当时还是美国斯坦福大学（Stanford University）研究生的佩奇（Larry Page）和布林（Sergey Brin）开始了对网页排序问题的研究。这两位小伙子之所以研究网页排序问题，一来是导师的建议（佩奇后来称该建议为"我有生以来得到过的最好建议"），二来则是因为他们对这一问题背后的数学产生了兴趣。

网页排序问题的背后有什么样的数学呢？这得从佩奇和布林看待这一问题的思路说起。

在佩奇和布林看来,网页的排序是不能靠每个网页自己来标榜的,无论把关键词重复多少次,垃圾网页依然是垃圾网页。那么,究竟什么才是网页排序的可靠依据呢?出身于书香门第的佩奇和布林(两人的父亲都是大学教授)想到了学术界评判学术论文重要性的通用方法,那就是看论文的引用次数。在互联网上,与论文的引用相类似的显然是网页的链接。因此,佩奇和布林萌生了一个网页排序的思路,那就是通过研究网页间的相互链接来确定排序。具体地说,一个网页被其他网页链接得越多,它的排序就应该越靠前。不仅如此,佩奇和布林还进一步提出,一个网页越是被排序靠前的网页所链接,它的排序就也应该越靠前。这一条的意义也是不言而喻的,就好比一篇论文被诺贝尔奖得主所引用,显然要比被普通研究者所引用更说明其价值。依照这个思路,网页排序问题就跟整个互联网的链接结构产生了关系,正是这一关系使它成为了一个不折不扣的数学问题。

思路虽然有了,具体计算却并非易事,因为按照这种思路,想要知道一个网页 $W_i$ 的排序,不仅要知道有多少网页链接了它,而且还得知道那些网页各自的排序——因为来自排序靠前网页的链接更有分量。但作为互联网大家庭的一员,$W_i$ 本身对其他网页的排序也是有贡献的,而且基于来自排序靠前网页的链接更有分量的原则,这种贡献与 $W_i$ 本身的排序也有关。这样一来,我们就陷入了一个“先有鸡还是先有蛋”的循环:要想知道 $W_i$ 的排序,就得知道与它链接的其他网页的排序,而要想知道那些网页的排序,却又首先得知道 $W_i$ 的排序。

为了打破这个循环,佩奇和布林采用了一个很巧妙的思路,即分析一个虚拟用户在互联网上的漫游过程。他们假定:虚拟用户一旦访问了一个网页后,下一步将有相同的几率访问被该网页所链接的任何一个其他网页。换句话说,如果网页 $W_i$ 有 $N_i$ 个对外链接,则虚拟用户在访问了 $W_i$ 之后,下一步点击那些链接当中的任何一个的几率均为 $1/N_i$。初看起来,这一假设并不合理,因为任何用户都有偏好,怎么可能以相同的几率访问一个网页的所有链接呢?但如果我们考虑到佩奇和布林的虚拟用户实际上是对互联网上全体用户

的一种平均意义上的代表，这条假设就不像初看起来那么不合理了。那么网页的排序由什么来决定呢？是由该用户在漫游了很长时间——理论上为无穷长时间——后访问各网页的几率分布来决定的，访问几率越大的网页排序就越靠前。

为了将这一分析数学化，我们用 $p_i(n)$ 表示虚拟用户在进行第 $n$ 次浏览时访问网页 $W_i$ 的几率。显然，上述假设可以表述为(请读者自行证明)：

$$p_i(n+1) = \sum_j \frac{p_j(n)p_{j\to i}}{N_j}$$

这里 $p_{j\to i}$ 是一个描述互联网链接结构的指标函数(indicator function)，其定义是：如果网页 $W_j$ 有链接指向网页 $W_i$，则 $p_{j\to i}$ 取值为 1，反之则为 0。显然，这条假设所体现的正是前面提到的佩奇和布林的排序原则，因为右端求和式的存在表明与 $W_i$ 有链接的所有网页 $W_j$ 都对 $W_i$ 的排名有贡献，而求和式中的每一项都正比于 $p_j$，则表明来自那些网页的贡献与它们的自身排序有关，自身排序越靠前(即 $p_j$ 越大)，贡献就越大。

为符号简洁起见，我们将虚拟用户第 $n$ 次浏览时访问各网页的几率合并为一个列向量 $\boldsymbol{p}_n$，它的第 $i$ 个分量为 $p_i(n)$，并引进一个只与互联网结构有关的矩阵 $\boldsymbol{H}$，它的第 $i$ 行 $j$ 列的矩阵元为 $H_{ij}=p_{j\to i}/N_j$，则上述公式可以改写为

$$\boldsymbol{p}_{n+1} = \boldsymbol{H}\boldsymbol{p}_n$$

这就是计算网页排序的公式。

熟悉随机过程理论的读者想必看出来了，上述公式描述的是一种马尔可夫过程(Markov process)，而且是其中最简单的一类，即所谓的平稳马尔可夫过程(stationary Markov process)[①]，而 $\boldsymbol{H}$ 则是描述马尔可夫过程中的转移概率分布的所谓转移矩阵(transition matrix)。不过普通马尔可夫过程中的转

---

① 马尔可夫过程，也称为马尔可夫链(Markov chain)，是一类离散随机过程，它的最大特点是每一步的转移概率分布都只与前一步有关。而平稳马尔可夫过程则是指转移概率分布与步数无关的马尔可夫过程(体现在我们的例子中，即 $\boldsymbol{H}$ 与 $n$ 无关)。另外要说明的是，本文在表述上不同于佩奇和布林的原始论文，后者并未使用诸如“马尔可夫过程”或“马尔可夫链”那样的术语，也并未直接运用这一领域内的数学定理。

移矩阵通常是随机矩阵(stochastic matrix),即每一列的矩阵元之和都为 1 的矩阵(请读者想一想,这一特点的“物理意义”是什么?)①。而我们的矩阵 $\boldsymbol{H}$ 却可能有一些列是零向量,从而矩阵元之和为 0,它们对应于那些没有对外链接的网页,即所谓的“悬挂网页”(dangling page)②。

上述公式的求解是简单得不能再简单的事情,即

$$\boldsymbol{p}_n = \boldsymbol{H}^n \boldsymbol{p}_0$$

其中 $\boldsymbol{p}_0$ 为虚拟读者初次浏览时访问各网页的几率分布(在佩奇和布林的原始论文中,这一几率分布被假定为是均匀分布)。

## 三、问题及解决

如前所述,佩奇和布林是用虚拟用户在经过很长——理论上为无穷长——时间的漫游后访问各网页的几率分布,即 $\lim_{n\to\infty}\boldsymbol{p}_n$,来确定网页排序的。这个定义要想管用,显然要解决三个问题:

(1) 极限 $\lim_{n\to\infty}\boldsymbol{p}_n$ 是否存在?

(2) 如果极限存在,它是否与 $\boldsymbol{p}_0$ 的选取无关?

(3) 如果极限存在,并且与 $\boldsymbol{p}_0$ 的选取无关,它作为网页排序的依据是否真的合理?

如果这三个问题的答案都是肯定的,那么网页排序问题就算解决了。反之,哪怕只有一个问题的答案是否定的,网页排序问题也就不能算是得到了满意解决。那么实际答案如何呢?很遗憾,是后一种,而且是其中最糟糕的情形,即三个问题的答案全都是否定的。这可以由一些简单的例子看出。比方

---

① 在更细致的分类中,这种每一列的矩阵元之和都为 1 的随机矩阵称为左随机矩阵(left stochastic matrix),以区别于每一行的矩阵元之和都等于 1 的所谓右随机矩阵(right stochastic matrix)。这两者在应用上基本是等价的,区别往往只在于约定。

② 这种几乎满足随机矩阵条件,但有些列(或行)的矩阵元之和小于 1 的矩阵也有一个名称,叫做亚随机矩阵(substochastic matrix)。

说，在只包含两个相互链接网页的迷你型互联网上，如果$\boldsymbol{p}_0=(1,0)^{\mathrm{T}}$，极限就不存在（因为几率分布将在$(1,0)^{\mathrm{T}}$和$(0,1)^{\mathrm{T}}$之间无穷振荡）。而存在几个互不连通（即互不链接）区域的互联网则会使极限——即便存在——与$\boldsymbol{p}_0$的选取有关（因为把$\boldsymbol{p}_0$选在不同区域内显然会导致不同极限）。至于极限存在，并且与$\boldsymbol{p}_0$的选取无关时它作为网页排序的依据是否真的合理的问题，虽然不是数学问题，答案却也是否定的，因为任何一个“悬挂网页”都能像黑洞一样，把其他网页的几率“吸收”到自己身上（因为虚拟用户一旦进入那样的网页，就会由于没有对外链接而永远停留在那里），这显然是不合理的。这种不合理效应是如此显著，以至于在一个连通性良好的互联网上，哪怕只有一个“悬挂网页”，也足以使整个互联网的网页排序失效，可谓是“一粒老鼠屎坏了一锅粥”。

为了解决这些问题，佩奇和布林对虚拟用户的行为进行了修正。首先，他们意识到无论真实用户还是虚拟用户，当他们访问到“悬挂网页”时，都不应该也不会“在一棵树上吊死”，而是会自行访问其他网页。对于真实用户来说，自行访问的网页显然与个人的兴趣有关，但对于在平均意义上代表真实用户的虚拟用户来说，佩奇和布林假定它将会在整个互联网上随机选取一个网页进行访问。用数学语言来说，这相当于是把$\boldsymbol{H}$的列向量中所有的零向量都换成$\boldsymbol{e}/N$（其中$\boldsymbol{e}$是所有分量都为1的列向量，$N$为互联网上的网页总数）。如果我们引进一个描述“悬挂网页”的指标向量（indicator vector）$\boldsymbol{a}$，它的第$i$个分量的取值视$W_i$是否为“悬挂网页”而定——如果是“悬挂网页”，取值为1，否则为0——并用$\boldsymbol{S}$表示修正后的矩阵，则

$$\boldsymbol{S}=\boldsymbol{H}+\frac{\boldsymbol{e}\boldsymbol{a}^{\mathrm{T}}}{N}$$

显然，这样定义的$\boldsymbol{S}$矩阵的每一列的矩阵元之和都是1，从而是一个不折不扣的随机矩阵。这一修正因此而被称为随机性修正（stochasticity adjustment）。这一修正相当于剔除了“悬挂网页”，从而可以给上述第三个问题带来肯定回答（当然，这一回答没有绝对标准，可以不断改进）。不过，这一修正解决不了前两个问题。为了解决那两个问题，佩奇和布林引进了第二个

修正。他们假定，虚拟用户虽然是虚拟的，但多少也有一些“性格”，不会完全受当前网页所限，死板地只访问其所提供的链接。具体地说，他们假定虚拟用户在每一步都有一个小于 1 的几率 $\alpha$ 访问当前网页所提供的链接，同时却也有一个几率 $1-\alpha$ 不受那些链接所限，随机访问互联网上的任何一个网站。用数学语言来说（请读者自行证明），这相当于是把上述 $\boldsymbol{S}$ 矩阵变成了一个新的矩阵 $\boldsymbol{G}$：

$$\boldsymbol{G} = \alpha \boldsymbol{S} + \frac{(1-\alpha)\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}}{N}$$

这个矩阵不仅是一个随机矩阵，而且由于第二项的加盟，它有了一个新的特点，即所有矩阵元都为正，（请读者想一想，这一特点的“物理意义”是什么？）这样的矩阵是所谓的素矩阵（primitive matrix）①。这一修正因此而被称为素性修正（primitivity adjustment）。

经过这两类修正，网页排序的计算方法就变成了

$$\boldsymbol{p}_n = \boldsymbol{G}^n \boldsymbol{p}_0$$

这个算法能给上述问题提供肯定答案吗？是的，它能。因为随机过程理论中有一个所谓的马尔可夫链基本定理（fundamental theorem of Markov chains），它表明在一个马尔可夫过程中，如果转移矩阵是素矩阵，那么上述前两个问题的答案就是肯定的。而随机性修正已经解决了上述第三个问题，因此所有问题就都解决了。如果我们用 $\boldsymbol{p}$ 表示 $\boldsymbol{p}_n$ 的极限②，则 $\boldsymbol{p}$ 给出的就是整个互联网的网页排序——它的每一个分量就是相应网页的访问几率，几率越大，排序就越靠前。

① 确切地说，这种所有矩阵元都为正的矩阵不仅是素矩阵，而且还是所谓的正矩阵（positive matrix）。这两者的区别是：正矩阵要求所有矩阵元都为正，而素矩阵只要求自己的某个正整数次幂为正矩阵。

② 读者们想必看出来了，$\boldsymbol{p}$ 其实是矩阵 $\boldsymbol{G}$ 的本征值为 1 的本征向量，而利用虚拟用户确定网页排序的思路其实是在用迭代法解决上述本征值问题。在数学上可以证明，上述本征向量是唯一的，而且 $\boldsymbol{G}$ 的其他本征值 $\lambda$ 全都满足 $\lambda<1$（更准确地说，是 $|\lambda|\leqslant\alpha$——这也正是下文即将提到的 $\boldsymbol{G}^n\boldsymbol{p}_0$ 的收敛速度与 $\alpha$ 有关的原因）。

这样，佩奇和布林就找到了一个不仅含义合理，而且数学上严谨的网页排序算法，他们把这个算法称为PageRank，不过要注意的是，虽然这个名称的直译恰好是“网页排序”，但它实际上指的是“佩奇排序”，因为其中的“Page”不是指网页，而是佩奇的名字。这个算法就是谷歌排序的数学基础，而其中的矩阵 $\boldsymbol{G}$ 则被称为谷歌矩阵（Google matrix）。

细心的读者可能注意到了，我们还遗漏了一样东西，那就是谷歌矩阵中描述虚拟用户“性格”的那个 $\alpha$ 参数。那个参数的数值是多少呢？从理论上讲，它应该来自于对真实用户平均行为的分析，不过实际上另有一个因素对它的选取产生了很大影响，那就是 $\boldsymbol{G}^n\boldsymbol{p}_0$ 收敛于 $\boldsymbol{p}$ 的快慢程度。由于 $\boldsymbol{G}$ 是一个 $N\times N$ 矩阵，而 $N$ 为互联网上——确切地说是被谷歌所收录的——网页的总数，在谷歌成立之初为几千万，目前为几百亿（并且还在持续增加），是一个极其巨大的数字。因此 $\boldsymbol{G}$ 是一个超大型矩阵，甚至很可能是人类有史以来处理过的最庞大的矩阵。对于这样的矩阵，$\boldsymbol{G}^n\boldsymbol{p}_0$ 收敛速度的快慢是关系到算法是否实用的重要因素，而这个因素恰恰与 $\alpha$ 有关。可以证明，$\alpha$ 越小，$\boldsymbol{G}^n\boldsymbol{p}_0$ 的收敛速度就越快。但 $\alpha$ 也不能太小，因为太小的话，“佩奇排序”中最精华的部分，即以网页间的彼此链接为基础的排序思路就被弱化了（因为这部分的贡献正比于 $\alpha$），这显然是得不偿失的。因此，在 $\alpha$ 的选取上有很多折中的考虑要做，佩奇和布林最终选择的数值是 $\alpha=0.85$。

以上就是谷歌背后最重要的数学奥秘。与以往那种凭借关键词出现次数所作的排序不同，这种由所有网页的相互链接所确定的排序是不那么容易做假的，因为作假者再是把自己的网页吹得天花乱坠，如果没有真正吸引人的内容，别人不链接它，一切就还是枉然①。而且“佩奇排序”还有一个重要特点，

① 当然，这绝不意味着在网页排序上已不可能再做假。相反，这种做假在互联网上依然比比皆是，比如许多广告或垃圾网页制造者用自动程序到各大论坛发帖，建立对自己网页的链接，以提高排序，就是一种常见的做假手法。为了遏制做假，谷歌采取了很多技术手段，并对有些做假网站采取了严厉的惩罚措施。这种惩罚（有时是误罚）对于某些靠互联网吃饭的公司有毁灭性的打击力。

那就是它只与互联网的结构有关，而与用户具体搜索的东西无关。这意味着排序计算可以单独进行，而无需在用户键入搜索指令后才临时进行。谷歌搜索的速度之所以快捷，在很大程度上得益于此。

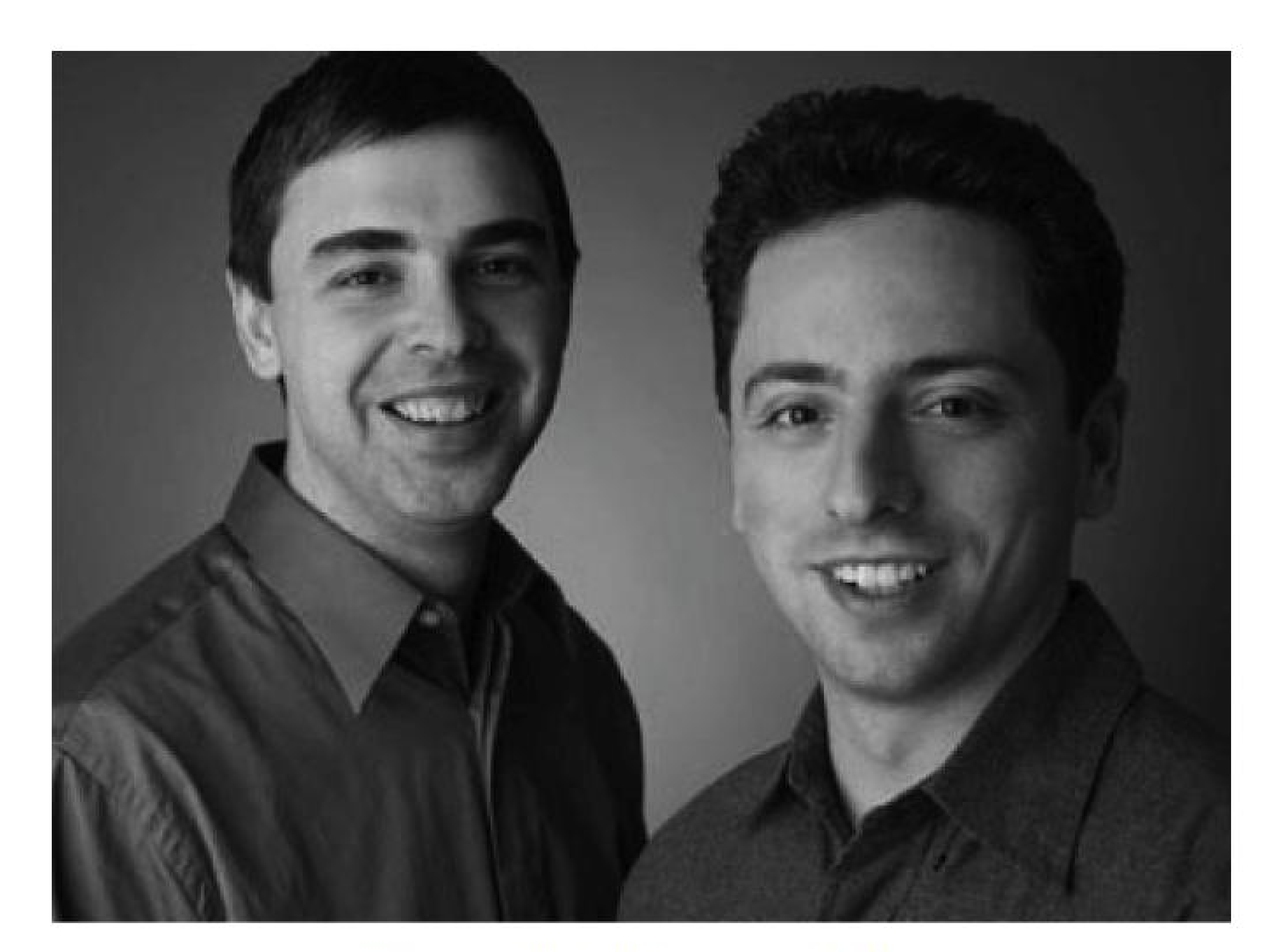
## 四、结语

在本文的最后，我们顺便介绍一点谷歌公司的历史。佩奇和布林对谷歌算法的研究由于需要收集和分析大量网页间的相互链接，从而离不开硬件支持。为此，早在研究阶段，他们就四处奔走，为自己的研究筹集资金和硬件。1998 年 9 月，他们为自己的试验系统注册了公司——即如今大名鼎鼎的谷歌公司。但这些行为虽然近乎于创业，他们两人当时却并无长期从商的兴趣。1999 年，当他们觉得打理公司干扰了自己的研究时，甚至萌生了卖掉公司的想法。

他们的开价是 100 万美元。

与谷歌在短短几年之后的惊人身价相比，那简直就是“跳楼大甩卖”。可惜当时却无人识货。佩奇和布林在硅谷“叫卖”了一圈，连一个买家都没找到。被他们找过的公司包括了当时搜索业巨头之一的 Excite（该公司后来想必连肠子都悔青了）。为了不让自己的心血荒废，佩奇和布林只得将公司继续办了下去，一直办到今天，这就是谷歌的“发家史”。

谷歌成立之初跟其他一些“发迹于地下室”（one-man-in-basement）的 IT 公司一样寒酸：雇员只有一位（两位老板不算），工作场所则是一位朋友的车库。但它出类拔萃的排序算法很快为它赢得了声誉。公司成立仅仅 3 个月，*PC Magzine* 杂志就把谷歌列为了年度最佳搜索引擎。2001 年，佩奇为“佩奇排序”申请到了专利，专利的发明人为佩奇，拥有者则是他和布林的母校斯坦福大学。2004 年 8 月，谷歌成为了一家初始市值约 17 亿美元的上市公司。不仅公司高管在一夜间成为了亿万富翁，就连当初给过他们几十美元“赞助费”的某些同事和朋友也得到了足够终身养老所用的股票回报。作为公司摇



谷歌公司创始人佩奇(左)和布林(右)

篮的斯坦福大学则因拥有“佩奇排序”的专利而获得了 180 万股谷歌股票。2005 年 12 月,斯坦福大学通过卖掉那些股票获得了 3.36 亿美元的巨额收益,成为美国高校因支持技术研发而获得的有史以来最巨额的收益之一①。

谷歌在短短数年间就横扫整个互联网,成为搜索引擎业的新一代霸主,佩奇和布林的那个排序算法无疑居功至伟,可以说,是数学成就了谷歌②。当然,这么多年过去了,谷歌作为 IT 界研发能力最强的公司之一,它的网页排

① 从投资角度讲,斯坦福大学显然是过早卖掉了股票,否则获利将更为丰厚。不过,这正是美国名校的一个可贵之处,它们虽擅长从支持技术研发中获利,却并不唯利是图。它们有自己的原则,那就是不能让商业利益干扰学术研究。为此,它们通常不愿长时间持有特定公司的股票,以免在无形中干扰与该公司存在竞争关系的学术研究的开展。

② 有些读者对“是数学成就了谷歌”这一说法不以为然,认为是佩奇和布林的商业才能,或将数学与商业结合起来的才能成就了谷歌。这是一个见仁见智的问题,看法不同不足为奇。我之所以认为是数学成就了谷歌,是因为谷歌当年胜过其他搜索引擎的地方只有算法。除算法外,佩奇和布林当年并无其他胜过竞争对手的手段,包括商业手段。如果让他们去当其他几家搜索引擎公司的老总,用那几家公司的算法,他们是不可能脱颖而出的;而反过来,如果让其他几家搜索引擎公司的老总来管理谷歌,用谷歌的算法,我相信谷歌依然能超越对手。因此,虽然谷歌后来确实用过不少出色的商业手段(任何一家那样巨型的公司都必然有商业手段上的成功之处),而当年那个算法在今天的谷歌——如正文所述——则早已被更复杂的算法所取代,但我认为谷歌制胜的根基和根源在于那个算法,而非商业手段,因此我说“是数学成就了谷歌”。

序方法早已有了巨大的改进，由当年单纯依靠“佩奇排序”演变为了由200多种来自不同渠道的信息——其中包括与网页访问量有关的统计数据——综合而成的更加可靠的方法。而当年曾给佩奇和布林带来过启示的学术界，则反过来从谷歌的成功中借鉴了经验，如今一些学术机构对论文影响因子（impact factor）的计算已采用了类似“佩奇排序”的算法。谷歌的发展极好地印证了培根（Francis Bacon）的一句名言：知识就是力量。

## 参考文献


[1] Austin D. How Google finds your needle in the Web's haystack[OL]. http://www.ams.org/samplings/feature-column/fcarc-pagerank.

[2] Battelle J. The birth of Google[J]. Wired, August 2005.

[3] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine[C]. Seventh International World-Wide Web Conference, Brisbane, Australia, April 14-18, 1998.

[4] Ibe O. Markov processes for stochastic modeling[M]. Amsterdam: Elsevier Academic Press, 2009.

[5] Langville A N, Meyer C D. Google's page rank and beyond: the Science of search engine rankings[M]. Princeton: Princeton University Press, 2006.

[6] Rousseau C, Saint-Aubin Y. Mathematics and technology[M]. Berlin: Springer, 2008.


2010年12月4日写于纽约

# 第二部分　物　　理



绘画：张京

# 从巴西的蝴蝶到得克萨斯的飓风[①]

## 一、决定论

在本书《时间旅行：科学还是幻想?》一文的第四节中，我们曾提到混沌理论中的一个概念：蝴蝶效应(butterfly effect)。这个效应也被称为对初始条件的敏感依赖性，指的是在某些——通常是非线性的——物理体系中，初始条件的细微改变有可能对体系的未来演化产生巨大影响。它的一种很富诗意的形容，是说巴西的一只蝴蝶拍动翅膀产生的空气扰动，有可能演变成美国得克萨斯州的一场飓风。这也是蝴蝶效应这一名称的主要由来。本文将对这一概念及其历史做一个简单介绍。

我们知道，人类描述自然的努力，很大程度上体现在对自然现象的时间演化进行描述上。这种描述在许多方面都取得了很大的成功。早在300多年前，英国科学家牛顿(Isaac Newton)就建立了我们称为牛顿力学的理论体系，对小至钟摆、陀螺，大至行星运动的各种自然现象的时间演化做出了极为精确的描述。1846年，天文学家们在牛顿力学所预言的位置近旁发现了几十亿千

① 本文的一个缩略修改版曾发表于《科幻世界》2007年第1期(科幻世界出版社出版)。

米之外的太阳系第 8 大行星——海王星，成为牛顿力学最辉煌的成就之一[①]。

牛顿力学的成功，除了体现在对某些自然现象时间演化的极为精确的描述外，还留下了一个非常重要的遗产，那就是决定论的思想。按照这一思想，从一个物理体系在某一时刻的状态，可以推算出它在任何其他时刻的状态。人们后来知道，牛顿力学本身只适用于描述一定范围内的力学现象，但它所留下的决定论思想却适用于几乎所有已知的物理定律，甚至在一定程度上包括了被公认为是非决定论性的量子力学[②]。

那么，决定论思想所具有的如此广泛的适用性，是否意味着我们在原则上可以对物理现象作出精确预言呢？在很长一段时间里，答案被认为是肯定的。但是，与这种被认为原则上可以做到的精确预言形成鲜明对比的，是实际上能精确求解的物理问题的稀少。以天体运动为例，人们能精确求解的只有二体问题。一旦把太阳、地球和月球这三个最熟悉的天体同时考虑进去，就没法精确求解了[③]。又比如流体运动，能精确求解的只有一些非常理想的情形，一旦把像黏滞性那样最常见的现实性质考虑进去，也就没法精确求解了。物理学家们能精确求解的问题，大都附加了各种简化条件。而真正的自然现象几乎从来都不满足那些条件，从而几乎没有一个是能精确求解的。

幸运的是，在那些无法精确求解的问题中，有一部分非常接近于某些能精确求解的问题。比如地球绕太阳的运转，所有其他天体的影响都相当微小，因此这一问题非常接近于能精确求解的二体问题。而且这两者的差异还可以通过各种手段加以弥补。正是由于这些近似手段——包括数值近似——的存在，使得物理学家们虽然很少能精确求解问题，却依然能对很多自然现象的演

① 不过后来的研究表明，海王星在距离理论预言非常近——相差不到 1°——的位置上被发现有一定的偶然性。关于这一点，可参阅拙作《那颗星星不在星图上：寻找太阳系的疆界》的第 20 章。

② 量子力学的状态演化是决定论性的，但量子测量过程是否也是决定论性的，则有一定的争议（虽然非决定论性的观点明显占优）。

③ 这还是在假定引力是由牛顿万有引力定律所描述的情况下，如果改用广义相对论，则连二体问题也无法严格求解。

化做出非常成功的描述。

## 二、早期研究

但是，任何近似手段都必然有误差，因此近似手段的有效性有赖于对误差的控制。随着研究的深入，物理学家们开始遇到了一些无法用近似手段来有效处理的问题。这些问题中有许多都具有蝴蝶效应，它使误差变得不可控制。19 世纪末，法国科学家庞加莱（Henri Poincaré）在对三体问题的研究中发现了一些这样的问题。他在《科学与方法》一书中写道："初始条件的微小差异有可能在最终的现象中导致巨大差异"，"预言变得不可能"。这或许是对蝴蝶效应最早的明确描述[①]。除三体问题外，流体力学中的湍流问题也是一种无法用近似手段来有效处理的问题。据说德国物理学家海森伯（Werner Heisenberg）曾经表示，有机会向上帝提问的话，他想问上帝为什么会有相对论？以及为什么会有湍流？他并且补充说："我确信上帝知道第一个问题的答案。"——言下之意是上帝也未必知道为什么会有湍流。

当科学家们接触到包含蝴蝶效应的问题或现象时，科幻小说家们也在用自己独特的方式描述着类似的现象。比如 1955 年，美国科幻小说家阿西莫夫（Isaac Asimov）写了一部小说，叫做《永恒的终结》（*The End of Eternity*）。在这部小说中，阿西莫夫描述了一群生活在物理时间之外的人，他们可以对人类历史进行修正，使其更加完美。但他们企图为人类创造一个完美历史的努力，在无形中扼杀了人类的创造与探索能力，使其在与外星生命的竞争中一败涂地。幸运的是，人类后来发现了这一点，并通过时间旅行的手段挽回了一切。在这部小说中阿西莫夫提到：对历史的每一次微小改变，都有可能以一种无法精确预言的方式改变数百万人的人生轨迹，这与蝴蝶效应的表述显然

① 不过《科学与方法》是一部科学哲学著作，庞加莱在自己的学术论文中并未明确表述过类似的结论。

有着极大的相似性。这种出现在科幻小说中的近乎先知先觉的描述，初看起来很令人吃惊，其实并不奇怪。因为现实世界本身就是一种最复杂的自然现象，像蝴蝶效应那样的东西，远在它成为科学研究的对象之前，就早已出现在了人们的日常经验中。所谓“差之毫厘，谬之千里”、“牵一发而动全身”等，都在一定程度上体现了这种效应。

但从那些日常体验上升为明确的理论表述，则是一个困难得多的问题。

从19世纪末到20世纪中叶，经过庞加莱、利雅普诺夫（Aleksandr Lyapunov）、弗兰克林（Philip Franklin）、马科夫（Andrei Andreevich Markov）、伯克霍夫（George David Birkhoff）等人的一系列研究，人们对这个困难得多的问题终于有了一定的认识。人们发现，对于满足一定条件的物理体系来说，只有周期性或近周期性（near periodic）的运动才不会因为初始条件的细微改变而产生剧烈变动。依照这个结果，如果运动是非周期性的，那么初始条件的细微改变就会对体系的演化造成巨大影响。因此，这个结果不仅确立了蝴蝶效应的存在，而且还对它的产生条件给出了一定的描述。但是，那时候人们最感兴趣的只是周期运动，因此有关非周期运动的结果虽可作为推论得到，在当时的学术文献中却极少提及。正因为如此，十几年后当美国科学家洛伦兹（Edward Norton Lorenz）在数值计算中再次遭遇蝴蝶效应的时候，依然感到了极大的惊讶。也正因为如此，发现蝴蝶效应的荣誉在很大程度上被后人归到了洛伦兹的头上。

## 三、模拟天气

洛伦兹是一位资深的气象学家，早在“二战”时期就在美国军方机构从事气象预测研究。战争结束后，洛伦兹来到了麻省理工学院（MIT），继续从事研究工作。从理论上预测气象变化——尤其是给出长期预测——是气象学家们梦寐以求的目标。但这一目标的实现却始终困难重重。这种困难是不难理解的，因为地球的大气层是一个巨大的流体系统，所有流体力学所具有的复杂

性，包括那个连上帝也未必知道起源的湍流问题，都会出现在大气层中。更何况，大气层的行为与海洋、地表、日照等各种复杂的外部条件都有着密切关系；而且大气层的组成相当复杂，其中有些组成部分——如水汽——的形态还常常会在气态、液态及固态之间变化。所有这一切，都使得气象预测成为了一个极其困难的课题。

在洛伦兹从事气象研究的时候，从理论上预测气象变化主要有两类方法。一类被称为动力气象学（dynamic meteorology），这类方法主要是把大气层看作一个流体系统，然后选取一些重要的物理量——如温度、风速等——进行研究。由于问题的高度复杂，人们还把大气层像切蛋糕一样分割成许多区域，每个区域都用一个点来代表。显然，这是极其粗糙的近似，但即便如此，整个大气层的状态往往还是需要几百万甚至更大数目的变量来描述①。换句话说，即便是求解一个非常粗糙的气候模型，往往也需要处理带有几百万个未知数的方程组。这无疑是极其困难的（但不是完全没有希望的）。除了动力气象学外，还有一类方法被称为天气学（synoptic meteorology），这类方法的特点是把对气候影响最大的一些大气结构，比如各种气旋，直接作为研究对象。天气学所使用的规律，有许多是描述那些大气结构的经验规律，而不是像流体力学那样系统性的物理理论。从这个意义上讲，天气学不如动力气象学那样基本。但天气学的优点，是把从动力气象学角度看非常复杂的某些大气结构作为了基本单元，从而有着独特的简化性。

洛伦兹所采用的主要是天气学方法。经过大量的简化后，洛伦兹得到了一个含有 14 个变量，且其中有一到两个变量的影响可以忽略的模型。但即使那样的模型用手工计算也是非常困难的，于是洛伦兹决定借助计算机的帮助。当时是 1959 年，距离个人计算机的问世还有二十几年。洛伦兹所使用的机器

① 举个例子来说，如果把大气层用长、宽、高分别为 100 千米、100 千米及 100 米的单元进行分割，则描述整个大气层——假定高度为 30 千米——的温度与风速所需的变量总数大约为 500 万。分割越细、引进的物理量越多，所需的变量数目也就越大。

用今天的标准来衡量是极为简陋的：体积庞大，噪声惊人，内存却只有今天普通个人计算机内存的百万分之一。经过几个月的努力（主要是编程），洛伦兹终于在那台机器上运行起了他的模拟天气。

## 四、奇怪的结果

日子平静地流逝着，洛伦兹与同事们间或地就模拟天气的演变打上一些小赌，聊以消遣。终于有一天，洛伦兹决定对某一部分计算进行更为仔细的分析。于是他从原先输出的计算结果中选出了一行数据——相当于某一天的天气状况——作为初始条件输入了程序。机器从那一天的数据开始了运行，洛伦兹则离开了办公室，去喝一杯悠闲的咖啡。中国的神话故事中有所谓“洞中方一日，世上已千年”的传说，洛伦兹的那杯咖啡就喝出了那样的境界。一个小时后，当他回到实验室时，他的模拟世界已经运行了两个月。洛伦兹一看结果，不禁吃了一惊！因为新的计算结果与原先的大相径庭。

这一结果为什么令人吃惊呢？因为这次计算所用的初始条件乃是从旧数据中选出来的。既然初始条件是旧的，所得的结果——在与旧数据可以比较的范围内——理应也跟旧数据相同，却怎么会大相径庭呢？洛伦兹的第一个反应是机器坏了——这在当时是经常发生的事情。但是，当他对结果做了更仔细的检验后，很快排除了那种可能性。因为他发现，新旧计算的结果虽然最终大相径庭，但在一开始却很相似，两者的偏差是在经过了一段指数增长过程之后才彻底破坏相似性的。如果机器坏了，是没有理由出现这种“有规律”的偏差的。

既然机器没有问题，那么究竟是什么原因造成了新旧计算之间的巨大偏差呢？洛伦兹很快找到了答案。原来，洛伦兹的程序在运行时保留了十几位有效数字，但在输出时为了让所有变量的数值能打印在同一行里，他对每个变量都只保留了小数点后 3 位有效数字。因此，当洛伦兹把以前输出的数据——即所谓旧数据——作为初始条件输入新一轮计算时，它与原先计算中

保留了十几位有效数字的数据相比,已经有了微小的偏差。洛伦兹的计算表明,在他的模拟系统中,这些微小的偏差每隔 4 天就会翻一番,直至新旧数据间的相似性完全丧失为止。

这正是蝴蝶效应!

由于蝴蝶效应的存在,洛伦兹意识到长期天气预报是注定不可能有高精度的。因为我们永远不可能得到绝对精确的初始条件,而且由于任何计算设备的内存都是有限的,我们在计算过程中也永远不可能保留无限的精度,所有这些误差都会因蝴蝶效应的存在而迅速(指数性地)扩大,从而不仅使一切高精度的长期气象预测成为泡影,而且葬送了建立在决定论思想之上的对物理现象进行精确预言的梦想①。

蝴蝶效应的发现还让洛伦兹回忆起一件他念本科时发生的事情。那是在 20 世纪 30 年代,当时他所在的镇上有许多学生迷上了弹球游戏(pinball game),那是一种让小球在一张插有许多小针的倾斜桌子上经过多次碰撞后进入特定小孔的游戏。当地政府曾想以禁止赌博为由禁止这种游戏,但游戏的支持者们争辩说这不是赌博,而是一种有关击球准确度的技巧比赛。他们的理由一度说服了政府官员,因为当时大家并不知道弹球游戏其实包含了蝴蝶效应,从而无论多高明的技巧都是无济于事的。

## 五、从蝴蝶到飓风

发现蝴蝶效应后的第二年,即 1960 年,洛伦兹在一次学术会议上粗略地提及了自己的发现,但没有发表详细结果。会议之后,洛伦兹感到自己的模型

---

① 严格地讲,由于无法得到精确的初始条件,以及无法在计算过程中保留无限的精度,即便没有蝴蝶效应,绝对精确的预言也是不可能的。但在没有蝴蝶效应的情况下,误差的影响往往是可控制的,蝴蝶效应的出现使误差的影响变得不可控制。另外需要说明的是,这里所说的“葬送了建立在决定论思想之上的对物理现象进行精确预言的梦想”与建立在微分方程解的存在及唯一性基础之上的决定论本身不是一回事,后者不会因为蝴蝶效应而破灭。

仍然太复杂，他决定寻找更简单的模型。1961 年，他从同事索兹曼(Barry Saltzman)那里得到了一个只含 7 个变量(即比他自己的模型少了一半的变量)的流体力学模型①。经过研究，洛伦兹很快发现，在索兹曼的模型中，有 4 个变量的数值很快就会变得可以忽略。因此，这一模型的真正行为可以用一个只含 3 个变量的方程组来描述，这个只含 3 个变量的方程组后来被冠上了洛伦兹的大名，称为洛伦兹方程组(Lorenz equations)。利用这一方程组，洛伦兹再次确认了蝴蝶效应的存在②，并于 1963 年在《大气科学杂志》(*Journal of the Atmospheric Sciences*)上发表了题为《确定性非周期流》(*Deterministic Nonperiodic Flow*)的论文，正式公布了自己的结果。

不过，无论是洛伦兹的原始论文，还是此后若干年内的其他有关著作，都没有直接使用“蝴蝶效应”这一名称。洛伦兹本人有时用海鸥造成的大气扰动来比喻初始条件的细微改变。“蝴蝶”这一“术语”的使用是在 9 年后的 1972 年。那一年洛伦兹要在华盛顿的一个学术会议上做报告，却没有及时提供报告的标题。于是会议组织者梅里利斯(Philip Merilees)“擅作主张”地替洛伦兹拟了一个题目：《巴西的蝴蝶拍动翅膀会引发德克萨斯的飓风吗？》(*Does the flap of a butterfly's wings in Brazil set off a tornado in Texas?*)。就这样，美丽的蝴蝶随着梅里利斯的想象飞进了科学术语之中③。

---

① 索兹曼与 20 世纪上半叶的那些科学家一样，对周期运动更感兴趣，因此没能在自己的模型上做出像洛伦兹那样的发现，虽然他在自己的模型中也已经发现了一些非周期性的解。

② 在这一点上，洛伦兹很受幸运女神的眷顾。他的方程组中含有一个被称为普朗特常数(Prandtl constant)的参数，这个参数对于水大约为 10，对于空气则大约为 1。洛伦兹与索兹曼都是气象学家，他们采用的数值原本应该是对应于空气的 1，但实际上两人却都采用了对应于水的 10。后来的研究发现，如果当时他们采用了对应于空气的普朗特常数，那个模型的解将是周期性的，洛伦兹将不可能得到他所需要的结果。

③ 不过那篇演讲的全文当时并未发表。另外需要提醒读者的是：蝴蝶效应的这一通俗表述有一定的误导性，容易让人以为在“蝴蝶拍动翅膀”与“得克萨斯的飓风”之间存在直接的因果联系。事实上，“蝴蝶拍动翅膀”和“得克萨斯的飓风”只是泛指初始条件的细微改变和体系未来演化的巨大变化，“得克萨斯的飓风”的物理起因有赖于无数的因素，绝非只是“蝴蝶拍动翅膀”。

除上述原因之外,“蝴蝶效应”的得名还有另外一个原因,那就是洛伦兹模型中有一个所谓的奇怪吸引子(strange attractor),它的形状从一定的角度看很像一只展翅的蝴蝶(图 1)。不过“蝴蝶效应”这一名称的最终风行,在很大程度上要归因于美国科普作家格雷克(James Gleick)的科普作品《混沌:开创新科学》(*Chaos: Making a New Science*)。这部被译成了多国文字,对混沌理论(蝴蝶效应是混沌理论的一部分)在世界范围内的热播起了极大促进作用的作品的第一章的标题就是《蝴蝶效应》。2004 年,蝴蝶效应甚至被搬上了银幕,成为一部科幻影片——虽然是不太成功的影片——的片名。



图 1 洛伦兹奇怪吸引子

蝴蝶效应及混沌理论在世界范围内的风行,一度使许多人产生一种错觉,以为物理学的又一次革命到来了。在这种“激情”的鼓舞下,这一领域涌现出了大量文章,其中包括不少低水平及浮夸的工作。从物理学的角度讲,蝴蝶效应及混沌理论并不包含新的原理性的东西,它们对物理学的最大启示是:形式上简单的物理学定律有可能包含巨大的复杂性,从而有可能解释比我们曾经以为的更为广阔的自然现象。这一点早在洛伦兹的论文发表之前,就已经被一些物理学家注意到了。20 世纪 60 年代初,美国物理学家费恩曼(Richard Feynman)在给本科生讲课——那些课程的内容后来汇集成了著名的《费恩曼物理学讲义》(*The Feynman Lectures on Physics*)——时,就非常清晰地阐述了这一点。他在介绍了流体力学中的若干复杂性之后这样写道:

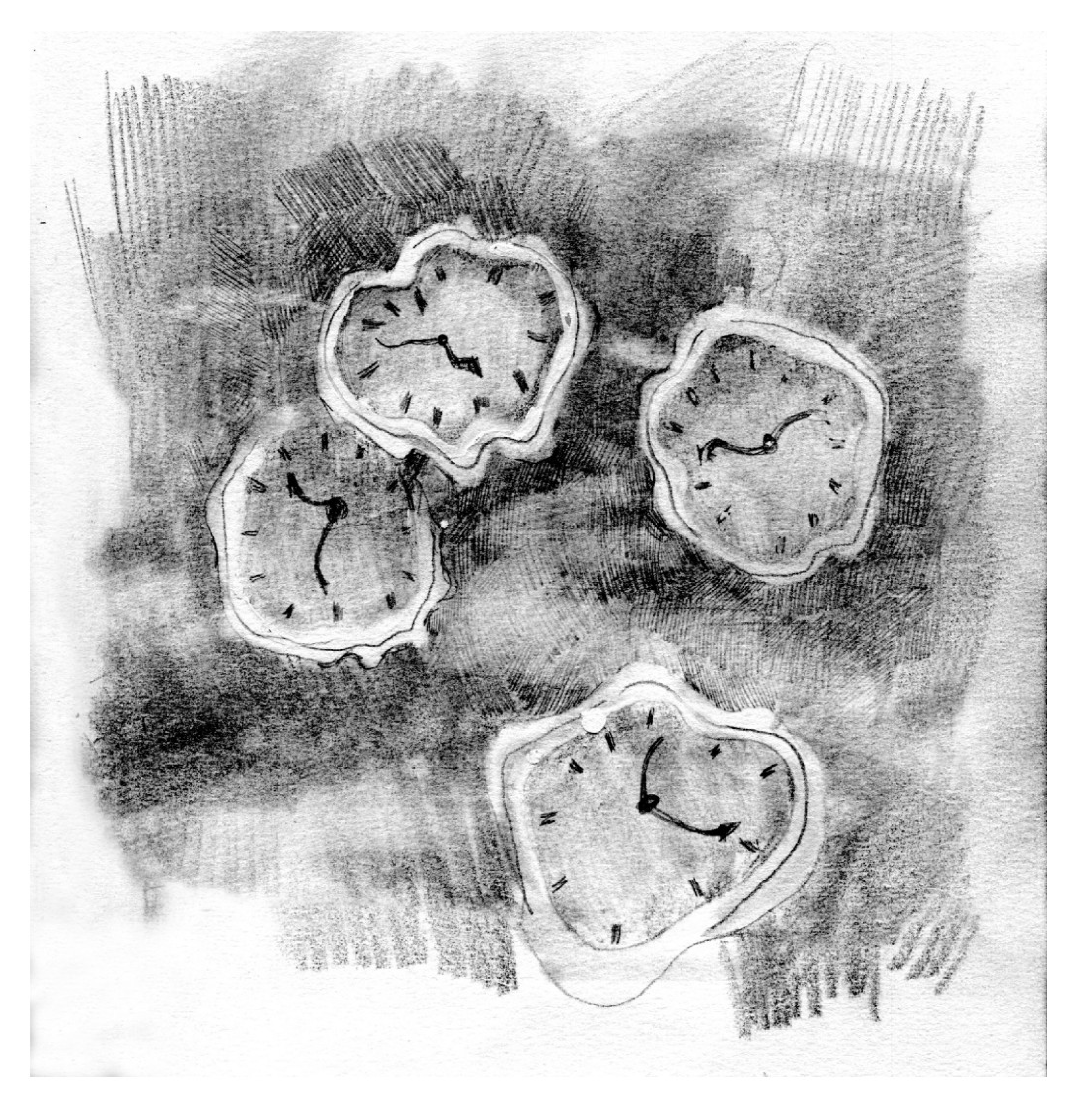
> 对物理学怀有莫名恐惧的人常常会说,你无法写下一个关于生命的方程式。嗯,也许我们能够。事实上,当我们写下量子力学方程式 $H\Psi=i\partial\Psi/\partial t$ 的时候,我们很可能就已在足够近似的意义上拥有了这样的方程式。我们刚才就看到了事物的复杂性可以多么容易且

富有戏剧性地逃脱描述它们的方程式的简单性。

费恩曼曾经希望人类的下一次智力启蒙会带给我们理解物理定律复杂内涵的方法。混沌理论的发展部分地体现了费恩曼的希望,但今天我们对这一领域的了解,在很大程度上依赖于计算技术的发展,与真正的智力启蒙还有一定的距离。真正的智力启蒙究竟会出现在什么时候?也许就像洛伦兹的天气一样,谁也无法准确预测,但我们会拭目以待。

2006 年 7 月 23 日写于纽约

2014 年 9 月 24 日最新修订



绘画：张京

# 关于时钟佯谬

## 一、时钟佯谬简史

在相对论的历史上，曾出现过一些流传很广的佯谬——也可以说是意外。之所以说是意外，是因为一些知名物理学家也参与了某些话题的讨论，给出的答案却不尽相同，从而使被讨论的话题变得更像佯谬。时钟佯谬（clock paradox）就是其中最著名的一个。

时钟佯谬源于一个很简单的问题：在**惯性参照系**中有两个彼此校准了的时钟，一个保持静止，另一个沿闭合路线运动后回到原地，问两个时钟重新相遇时哪个时钟慢了？

最早对这个问题作出回答的当然不是别人，而是爱因斯坦（Albert Einstein）本人。他在狭义相对论的开山之作《论动体的电动力学》（*On the Electrodynamics of Moving Bodies*）中对这一问题给予了明确回答，答案是运动时钟慢了，理由是狭义相对论的时钟延缓（time retardation）效应①。但不

① 也称为时间膨胀（time dilation）效应。

知是没往那个角度考虑，还是觉得那不是问题，爱因斯坦只在静止时钟参照系中回答了这一问题，而未如许多后人所做的那样，将静止时钟参照系与运动时钟参照系视为对等来进行分析，从而没有触及后来被称为“佯谬”的东西——即两个时钟均认为自己静止，对方运动，从而在表观上彼此矛盾地均认为是对方慢了。

文献中最早触及时钟佯谬，并试图给予解释的知名人士是法国物理学家朗之万(Paul Langevin)。他在 1911 年发表的一篇题为《空间和时间的演进》(*The Evolution of Space and Time*)的文章中提出：解释时钟佯谬的要点，是注意到运动时钟经历了加速运动，而静止时钟没有经历加速运动。在那篇文章中，他还引入了两条很受后人欢迎的“人性化措施”：一是将时钟换成人，二是采用两人互发光信号的方法来比较各自经历的时间。朗之万这篇文章影响了很多人，时钟佯谬的别名“双生子佯谬”(twin paradox)据说就是他将时钟换成人所引发的(图 2)。直至今日，许多初等教材及科普读物仍采用朗之万的方法分析时钟佯谬(有些著作还将两人互发光信号改成更“亲密”地互数对方心跳)。

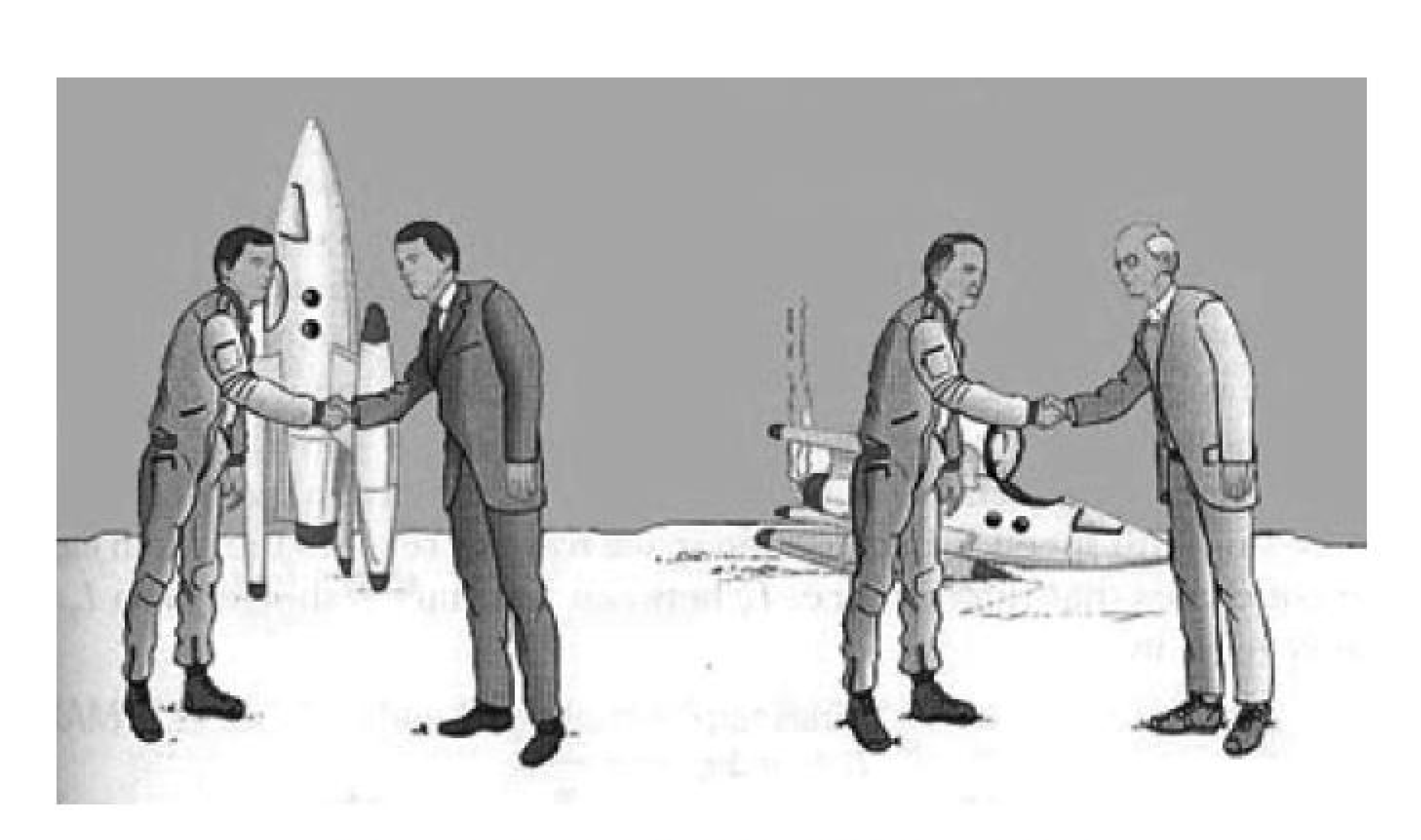


图 2　时钟佯谬的别名是双生子佯谬

继朗之万之后，德国物理学家冯·劳厄(Max von Laue)也对时钟佯谬提出了解释。他在 1912 年发表的一篇题为《两种反相对论意见及对它们的反驳》(*Two Objections against the Theory of Relativity and Their Refutation*)的文章中提出：解释时钟佯谬的要点，是注意到运动时钟在从远

离转为返回的过程中更换了参照系，而静止时钟没有更换参照系。他的这一看法也被一些人采纳，成为分析时钟佯谬的切入点之一。

爱因斯坦本人则似乎直到广义相对论发表之后，才对时钟佯谬中的“佯谬”部分发表看法。他在1918年发表的一篇题为《关于反相对论意见的对话》(*Dialog about Objections against the Theory of Relativity*)的文章中运用自己的“独门武器”等效原理提出了一种看法。他认为解释时钟佯谬的要点，是注意到运动时钟受到了与加速场等效的引力场的影响，而静止时钟没有受到那样的影响。由于引力场中的时钟延缓是广义相对论的推论，加上爱因斯坦在相对论领域的巨大威望，他的解释在很长的时间里被视为了时钟佯谬的正解。20世纪上半叶的很多相对论名著，比如奥地利物理学家泡利(Wolfgang Pauli)的《相对论》(*Theory of Relativity*)、丹麦物理学家莫勒(*Christian Møller*)的《相对论》(*Theory of Relativity*)、美国物理学家托曼(Richard C. Tolman)的《相对论、热力学及宇宙学》(*Relativity, Thermodynamics, and Cosmology*)等都采用了爱因斯坦的观点，认为时钟佯谬需要用广义相对论来解释。甚至连后来出版的一些知名著作，比如20世纪70年代出版的日本物理学家汤川秀树的《经典物理学》，也认为时钟佯谬“在狭义相对论框架中考虑时是佯谬，但若考虑广义相对论就不再是佯谬了”。

不过，以上这些物理学家对时钟佯谬的解释虽各有各的侧重点，而且从现代观点来看，都不够切中要害，但他们的结论是一致的，并且也是正确的，那就是运动时钟慢了(这一结论后来得到了实验证实)。时钟佯谬作为“佯谬”的正史大体就是这些，但在结束本节之前，有一段“外史”必须提一下，因为时钟佯谬作为“佯谬”的名声，在很大程度上其实是被那段“外史”搅起来的。那段“外史”就是英国哲学家丁格尔(Herbert Dingle)在20世纪50年代末对相对论的猛烈攻击，而那攻击的主要目标就是时钟佯谬。丁格尔在攻击中先是认为两个时钟应显示相同时间，遭驳斥后又转而宣称相对论的预言与经验不符(实际上有关时钟佯谬的预言当时就已得到了$\mu$子衰变实验的支持)，甚至连数学上显而易见的洛伦兹变换存在逆变换这一基本事实，也被他斥为明显不可能。

这样一个用上海话讲根本就“拎勿清爽”的人物照说是不配在本节中被提及的，但历史有时是充满惊奇的，这位“拎勿清爽”的丁格尔先生在1951—1953年间竟担任过英国皇家天文学会（Royal Astronomical Society）的主席①，并且还是英国科学史学会（British Society for the History of Science）的创始成员之一以及1955—1957年间的主席。也许是因为这些背景的缘故，几十位物理学家对他那破绽百出的文字进行了认真驳斥，从而构成了时钟佯谬历史上一段虽无价值，却引人注目的“外史”，并在最大程度上成就了时钟佯谬作为“佯谬”的名声。

## 二、时钟佯谬简析

以上就是时钟佯谬的简史，对于我的读者来说，还可以补充一段史上最小的八卦，那就是我“小时候”也曾认同过广义相对论才是时钟佯谬正解的看法，在自己网站（http://www.changhai.org/）的昔日版本中还贴过怀旧之作，对某些基于朗之万和冯·劳厄思路的解释进行了“呛声”。也许是因此之故，常有网友问及时钟佯谬。从这个意义上讲，本文可算是一篇还债之作——还那怀旧之作引发的文债。

这篇还债之作之所以拖到今天才写，不是企图“赖债”，而是因为时钟佯谬的现代解释实在太简单了，简直就是“一句话解释”，就算加上注解，似乎也构不成一篇文章。当然，这种估计如今看来显然是错误的，因为真要写的话，几乎没什么话题是构不成文章的。

好了，现在言归正传，谈谈时钟佯谬的现代解释。自20世纪70年代以来，有关相对论的许多教材和专著，比如沃尔德（Robert M. Wald）的《广义相

① 我试图查找丁格尔对天文学的贡献，却没能找到。他最主要的天文活动似乎是参加了1927年与1932年的日食远征队，但两次都因天气原因无功而返。他被选为皇家天文学会主席一事，据说连他自己都觉得惊讶，因为自20世纪30年代后期起，他就已经离开天文学，转而研究自然哲学了。

对论》(*General Relativity*)、托雷提(Roberto Torretti)的《相对论与几何》(*Relativity and Geometry*)、伦德勒(Wolfgang Rindler)的《相对论》(*Relativity*)、萨克斯(Rainer Sachs)等人的《数学家用广义相对论》(*General Relativity for Mathematicians*)、米斯纳(Charles W. Misner)等人的《引力》(*Gravitation*)、塞克斯尔(Roman U. Sexl)等人的《相对论、群论、粒子》(*Relativity, Groups, Particles*)等都采用了几何语言来阐述时钟佯谬,这就是所谓时钟佯谬的现代解释。在中文文献中,梁灿彬等人的《微分几何入门与广义相对论》也采用了现代解释,中文读者可以参考①。

那么究竟什么是时钟佯谬的现代解释呢?我没有忽悠诸位,它的要点确实只有一句话,那就是:**时钟记录的是自己的世界线长度**。在时钟佯谬中,之所以是运动时钟慢了,原因就是它的世界线长度较短。这里唯一要说明的是,所谓“世界线长度”指的是闵科夫斯基空间中的长度②,它与普通空间中的长度有一个最大的区别,那就是前者的测地线(即“直线”)长度是极大值而不是极小值。(请读者想一想,这是闵科夫斯基空间的什么特点造成的?)但无论闵科夫斯基空间还是普通空间,有一点是共同的,那就是长度是坐标变换下的不变量,从而与参照系或坐标系的选择无关③。

在这样的几何语言下,**时钟佯谬的结论,即运动时钟比静止时钟慢,不过**

---

① 不过,梁灿彬等人的著作多加了一个似是而非的论据,即认为三维加速度是相对的,四维加速度才是绝对的,以此反驳那种认为加速度也是相对的观点。其实,**就该书所述的情形——即该书自己援引的第6.3节——而言**,在对解释时钟佯谬来说最关键的**加速度的“有”和“无”的区分上**,三维加速度与四维加速度都是绝对的(理由很简单,相对于一个惯性系作加速运动的物体相对于任何惯性系都是作加速运动的,从四维加速度的分量表达式也可看出,四维加速度为零当且仅当三维加速度为零),对两者作相对与绝对的划分**对于解释时钟佯谬来说**不仅似是而非,而且毫无必要。

② 对于类空曲线来说,这种长度常被称为“固有长度”(proper length),对于类时曲线来说,则常被称为“原时”或“固有时”(proper time)。另外,闵科夫斯基空间常被称为“闵科夫斯基时空”。

③ 本文对“参照系”和“坐标系”这两个术语只作粗略区分:意在强调与核心物理观察者(即那两个时钟或双生子中的某一个)的关系时用“参照系”,意在强调具体数学坐标时用“坐标系”。

**是对两个时钟的世界线长度不同这一简单事实的简单陈述而已**,并不比普通空间中两条曲线的长度不同来得奥妙,更没有任何佯谬可言。这一点是如此的显而易见,以至于前面提到的《相对论与几何》一书的作者托雷提感慨道:“相对论时钟是类时世界线上的里程表,假如人们对这一事实有过更多关注,那么在所谓时钟佯谬上付出过的很多努力就可以省掉了。”

但话虽如此,如果我们在这里就结束本文,有些读者也许会感到失望,因为在时钟佯谬的传统讨论中,人们曾花大力气讨论运动时钟参照系,试图说明该参照系也能理解运动时钟变慢这一结论。即便那些努力如今“可以省掉了”,但若不把那最令人困惑的运动时钟参照系单独拿出来,更直接地讨论一下,似乎多少有些偷懒的感觉。为了“抚平”这种感觉,我们再多说几句。

如前所述,在时钟佯谬的现代解释中,运动时钟之所以慢了,原因是它的世界线长度较短。如果画出时空图的话,静止时钟的世界线是直线,运动时钟的世界线是曲线(参阅图 3),两者起始点相同,但曲线的长度较短(因为是闵科夫斯基空间)。这一切当然都是几何语言。那么,在这种语言中运动时钟参照系是什么呢?它就是把运动时钟的世界线视为直线,而把静止时钟的世界线视为曲线的坐标系。这种坐标系其实我们并不陌生,它就是曲线坐标系——把运动时钟的世界线作为时间轴的曲线坐标系。明白了这一点,运动时钟参照系里的问题就迎刃而解了,因为曲线坐标系虽然完全合法,而且确实能在表观上使两条世界线的“曲”、“直”互换,却不会改变它们的长度,从而不会改变时钟佯谬的结论,因为曲线坐标系有一个众所周知的“副作用”,那就是会改变度规的形式,使之不再是闵科夫斯基度规 $ds^2 = \eta_{\mu\nu} dx^\mu dx^\nu$ 或欧几里得度规 $ds^2 = \delta_{ij} dx^i dx^j$。比如极坐标下的度规是 $ds^2 = dr^2 + r^2 d\theta^2$ 而不是 $ds^2 = dr^2 + d\theta^2$。正是这种度规改变抵消了“曲”、“直”互换

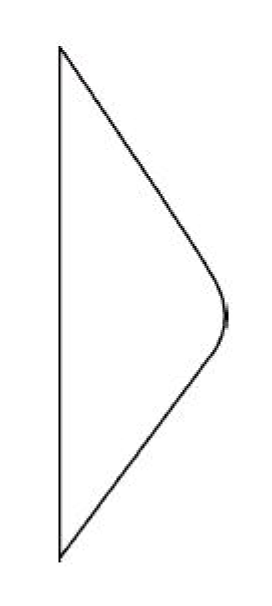


图 3 时钟佯谬的时空图

的影响，使得长度不变，从而保证了时钟佯谬的结论不变①。

当然，这一切其实就是对“长度是坐标变换下的不变量”这一简单事实的繁琐说明，只不过这样一说明，或许显得更像是“解释”而已。另外，它也示范了一种方法，即当我们对时钟佯谬的某个方面感到困惑时，想想它在几何语言下的对应，以及在普通空间中的类比，往往会豁然开朗。

在本节的最后，我们评论一下“时钟佯谬需要用广义相对论来解释”这一流传很广的观点。很明显，时钟佯谬的现代解释并不支持这种观点。**时钟佯谬作为闵科夫斯基空间中的现象，是完全可以，并且也应该用狭义相对论来解释的**——正如上述现代解释所做的那样。事实上，在闵科夫斯基空间中无论采用什么参照系或坐标系，都不可能使四维曲率张量非零，从而不可能出现曲率意义下的引力场。不仅如此，**迄今为止除上述现代解释外，对时钟佯谬的任何其他解释都是针对特例或近似的**。比如朗之万和冯·劳厄的解释通常只被用于运动时钟匀速远离，再匀速飞回的特例；爱因斯坦的解释则往往要采用

---

① 在讨论本文的过程中，有网友提出了这样一个问题：为什么运动时钟参照系必须接受一个“不平等”的度规，而不能像静止时钟参照系那样，认为自己的度规是闵科夫斯基度规？在时钟佯谬的框架中，这是因为一开始就已假定问题发生在闵科夫斯基空间中，而所谓“静止”时钟与“运动”时钟的唯一合理的定义就是前者的世界线为测地线，后者的世界线为非测地线，而且两者都是——并且也只能是——相对于背景度规来定义的（相对论不是一个马赫式的理论，在相对论中与奥地利哲学家马赫所设想的遥远星体所起作用最接近的东西就是背景度规），这就保证了只有前者所在的参照系可以自始至终使用闵科夫斯基度规，后者则只能使用从闵科夫斯基度规（通过坐标变换）诱导出来的度规。

不过，这也引出了一个更一般的问题，那就是闵科夫斯基度规的特殊地位是从何而来的？在狭义相对论中，这可以说是一个基本假设（或经验事实）。那么，广义相对论的情况是否会强一些呢？它是否能对闵科夫斯基度规的特殊地位做出“更物理”的说明（从而也对时钟佯谬作出“更物理”的解释）呢？很遗憾，答案是否定的，因为闵科夫斯基度规的特殊地位在广义相对论中也是基本假设，因为广义相对论所用的赝黎曼空间就是局部为闵科夫斯基空间的流形（这是等效原理的体现），其度规则是可以局部地由闵科夫斯基度规诱导出来的。实际上，按照我们在正文中所建议的类比思路，闵科夫斯基度规在相对论中的地位与欧几里得度规在普通黎曼几何中的地位是完全相似的，两者都是切空间中的度规，都是诱导其他度规的基石。广义相对论无法比狭义相对论“更物理”地解释闵科夫斯基度规的特殊地位（从而也无法“更物理”地解释时钟佯谬），就好比黎曼几何无法比欧几里得几何更充分地说明欧几里得度规的特殊地位。

广义相对论的弱场近似。与之相比,时钟佯谬的现代解释完全不受那些特例或近似的约束,从而有极大的普适性。哪怕两个时钟都作任意复杂的类时运动,现代解释依然适用(传统解释则会变得苦不堪言)。甚至当我们把时钟佯谬的舞台由闵科夫斯基空间搬到更复杂的空间,从而越出狭义相对论的范围时,现代解释依然适用(只需增添一个非平凡的背景度规即可)①。

## 三、关于理想时钟

在结束本文前,我们还要讨论一个衍生话题:什么是时钟?之所以要讨论这个话题,是因为时钟佯谬的传统解释,尤其是爱因斯坦的思路,很容易产生一个与"什么是时钟?"密切相关的问题,那就是加速度究竟会不会对时钟产生影响?关于这个问题,许多现代教材及专著——比如前面提到的托雷提、塞克斯尔、伦德勒、米斯纳等人的著作——都给出了明确回答,我们在这里作一个简单介绍。

首先要指出的是,对于具体的时钟来说,这个问题的答案显然与时钟的结构有关(而且大都是肯定的),比如对加速场中的摆钟来说,加速度越大,摆动的周期就越短,如果我们用这种摆钟的摆动次数来计时,加速度对它显然是有影响的。又比如对人来说,如果我们将生理节律作为时钟——就像郎之万所做的那样,它显然也会受加速度影响,在足够大的加速度下——对飞行员来说是 $10g$ 以上,对本文作者来说估计 $5g$ 就够了——甚至会"停止计时"(一命呜呼)。不仅宏观世界的时钟如此,曾被用来验证相对论的原子钟,严格讲也是会受加速度影响的,因为它的能级结构与包括加速场在内的各种外场有关。甚至连最早对时钟延缓效应作出实验判决的 $\mu$ 子的衰变,我们也并不肯定它

① 有人也许要问:时钟佯谬的传统解释到底算不算错误?我的看法是,在各自针对的特例或近似下,它们作为理解时钟佯谬的辅助手段,谈不上错误。但它们是否称得上解释,则取决于对"解释"一词的理解,我个人认为它们起码不算是好的解释。

不会受加速度影响。只不过，对于微观世界的时钟来说，与它内部的微观相互作用相比，加速场的影响往往是微乎其微的，因此当我们采用微观世界的时钟时，通常都能忽略加速度的影响。

但无论加速度对具体时钟的影响是有还是无，是大还是小，有一点是肯定的，那就是我们并不认为像摆钟受加速度影响，或本文作者在 $5g$ 的加速度下"停止计时"那样的效应反映了时间的固有性质。相反，我们认为那是具体时钟的缺陷导致的表观效应，是可以、并且必须校正的。我们真正关心的是反映时间本质的时钟，即所谓的理想时钟。本文所说的时钟除非有特别说明，指的也全都是理想时钟。

因此我们的问题其实是：什么是理想时钟？对此，相对论——无论狭义相对论还是广义相对论——的回答是：理想时钟是记录自己世界线长度的时钟。这是理想时钟的定义，被托雷提称为"时钟假设"(clock hypothesis)。不难证明，对于时钟佯谬所涉及的闵科夫斯基空间的时钟来说，这一定义给出的理想时钟与瞬时随动惯性系(momentarily co-moving inertial frame)里的时钟完全同步(请读者自行证明)[①]。由此，我们也得到了"加速度究竟会不会对时钟产生影响?"的答案，那就是加速度对理想时钟没有影响。

细心的读者也许已经注意到了，上述理想时钟的定义其实正是前面提到过的时钟佯谬现代解释的要点，即"时钟记录的是自己的世界线长度"。时钟佯谬的现代解释之所以有极大的普适性，一个很根本的原因就是它实际上包含了理想时钟的定义。

在本文的最后，给感兴趣的读者留两组思考题：

(1) 人们常说的"引力场中的时钟较慢"究竟是什么意思？把它与等效原理合在一起，是否会得出与"加速度对理想时钟没有影响"相矛盾的结论？

(2) 在理想时钟的定义中，只校正了加速度的影响，这是否是一种随意选

---

① 瞬时随动惯性系是指在所考虑的时刻与运动时钟具有相同瞬时速度的惯性参照系，也称为"瞬时静止惯性系"(momentary inertial rest frame)。

择？能否把速度的影响也像加速度的影响一样校正掉？

## 参考文献

［1］ Misner C W，et al. Gravitation［M］. New York：W. H. Freeman，1973.

［2］ Rindler W. Relativity：special，general，and cosmological［M］. Oxford：Oxford University Press，2006.

［3］ Sachs R K，Wu H H. General relativity for mathematicians［M］. Berlin：Springer，1983.

［4］ Sexl R U，et al. Relativity，groups，particles：special relativity and relativistic symmetry in field and particle physics［M］. Berlin：Springer-Verlag，2001.

［5］ Torretti R. Relativity and geometry［M］. NewYork：Dover Publications，1996.

［6］ 梁灿彬，周彬. 微分几何入门与广义相对论（上册）［M］. 北京：科学出版社，2006.

2011 年 5 月 14 日写于纽约

# 从等效原理到爱因斯坦-嘉当理论

## 一、等效原理

众所周知，等效原理(equivalence principle)——即引力场与加速场的不可区分性——是局域的。在一个非局域的参照系——比如有限大小的“爱因斯坦升降机”(Einstein’s elevator)——中，我们可以通过对所谓“测地偏离”(geodesic deviation)效应的观测，来区分引力场与加速场。这种观测之所以有效，是因为所涉及的是联络(connection)的导数，或者说曲率(curvature)的分量，这是不能通过等效原理消去的。由于对测地偏离效应的观测是在有限大小而非局域的参照系中进行的，因此与等效原理并不矛盾。

一般教材的讨论大都到此为止。

很明显，若所有物理效应都只跟度规及联络有关，那等效原理的成立就是普遍的。但假如存在某种局域的物理效应与曲率相耦合，那么哪怕在局域的参照系中，我们也将可以通过对这种物理效应的观测，而对引力场与加速场做出区分。

那样的物理效应是否存在呢？答案极有可能是肯定的。事实上，有自旋

粒子的运动很可能就是那样的物理效应之一。虽然迄今尚无任何实验足以检验这类效应，但一般认为，有自旋粒子在引力场中的运动由所谓的“马西森-帕帕佩特鲁-狄克逊方程”（Mathisson-Papapetrou-Dixone quation）所描述，而这一方程显含曲率张量。因此，有自旋粒子在引力场中的运动会与曲率相耦合。由此得出的一个推论则是，通过观测有自旋粒子的运动，原则上能在局域参照系中区分引力场与加速场①。

从某种意义上讲，这意味着等效原理不再成立了。

但是，这并不意味着广义相对论失效。对于广义相对论来说，等效原理的作用主要是确立时空的赝黎曼（pseudo-Riemannian）结构。为此只要在每一点上存在局域参照系，使度规为闵科夫斯基度规（Minkowski metric），同时使得联络系数全部为零即可（如果把这作为等效原理的定义，则等效原理的成立将不受上面提到的效应所影响）。至于是否有物理效应与曲率相耦合，并不妨碍广义相对论的建立。有自旋粒子的经典运动在广义相对论的框架中是完全可以处理的，就像时钟佯谬在狭义相对论的框架中完全可以处理一样。

## 二、爱因斯坦-嘉当理论

刚才我们提到，有自旋粒子在引力场中的运动会与曲率相耦合，从而能用来局域地区分引力场与加速场。这一讨论只涵盖了与引力有关的有自旋粒子问题的一半——即有自旋粒子在给定的引力场中会如何运动。现在，我们来考虑问题的另一半，即有自旋粒子本身会产生什么样的引力场。这是一个性质很不相同的问题，因为有自旋粒子在给定的引力场中的运动——如前所述——不会对广义相对论的结构产生根本性的影响，而有自旋粒子本身产生

① 这里需要注意的是，所谓“有自旋粒子”指的是量子场论意义下的有自旋的点粒子，因为这里所借重的是量子场论意义下的“自旋”和“点粒子”这两个概念——假如所讨论的不是这种概念，而是有限大小的经典旋转物体，则与等效原理的成立与否无关（因为它不是局域的）。从某种意义上讲，这是在通过量子效应来局域地区分引力场与加速场。

的引力场，则——如我们即将看到的——虽非必然，却很有可能把我们引向不同于广义相对论的理论，比如爱因斯坦-嘉当(Einstein-Cartan)理论。

我们知道，对所有具有能量动量起源的角动量 $J^{abc}=x^aT^{bc}-x^bT^{ac}$ 来说，能量动量张量 $T^{ab}$ 的守恒(即$\partial_aT^{ab}$)与对称(即 $T^{ab}=T^{ba}$)保证了角动量的守恒(即$\partial_aJ^{abc}=0$)。这种角动量被称为轨道角动量，它涵盖所有的经典角动量(包括经典意义下的"自旋"——即自转角动量)。另一方面，我们也知道，并非所有的角动量都具有能量动量起源，比如量子意义下的自旋就不具有能量动量起源(因为一个有自旋粒子完全可以是无质量的)。如果我们把这种所谓"内禀"(即不具有能量动量起源)的角动量记为 $S^{abc}$，则总角动量可以表示为 $J^{abc}=S^{abc}+x^aT^{bc}-x^bT^{ac}$。这时角动量守恒$\partial_aJ^{abc}=0$ 将会要求

$$\partial_aS^{abc}=T^{cb}-T^{bc}$$

这一式子表明，除非内禀角动量单独守恒(即$\partial_aS^{abc}=0$)，否则能量动量张量将是非对称的(即 $T^{ab}\neq T^{ba}$)。由于内禀角动量显然并不单独守恒，因此上式中的能量动量张量是非对称的。

如果能量动量张量非对称，那么爱因斯坦场方程 $G^{ab}=8\pi T^{ab}$ 将要求爱因斯坦张量 $G^{ab}$ 也是非对称的。这表明时空几何将不会是单纯的黎曼几何(Riemannian geometry)。使 $G^{ab}$ 非对称的一种最简单的方案，就是引进非零的时空挠率(torsion)$t^a_{bc}=\Gamma^a_{bc}-\Gamma^a_{cb}$。由此产生的最简单的理论就是所谓的爱因斯坦-嘉当理论，是法国数学家嘉当(Élie Cartan)于 1922 年提出的。

与纯度规性的广义相对论不同，爱因斯坦-嘉当理论是一种建立在仿射联络(affine connection)基础上的引力理论，在这种理论中等效原理不再成立(因为非零挠率使得联络系数全部为零的局域参照系不复存在)。爱因斯坦-嘉当理论中的这种带挠率的几何被称为黎曼-嘉当几何(Riemann-Cartan geometry)。爱因斯坦-嘉当理论的场方程则为

$$G^{ab}=8\pi T^{ab}$$

$$t^a_{bc}=8\pi S^a_{bc}+4\pi\delta^a_bS^d_{cd}+4\pi\delta^a_cS^d_{db}$$

不过，上述推理并不是唯一的。

这不仅是因为使能量动量张量非对称的方法并不唯一(从而爱因斯坦-嘉当理论并不是唯一可能的推广),而且也是因为内禀角动量的出现及并不单独守恒这一特点并非必然导致能量动量张量的非对称性。事实上,通过对能量动量张量添加一个对运动方程没有影响的散度项,我们总可以将它改写为对称形式。这种对称形式的能量动量张量被称为贝林番特张量(Belinfante tensor)。有一种(比较常见的)观点认为,出现在爱因斯坦场方程中的能量动量张量应该是贝林番特张量①。显然,这可以使得爱因斯坦场方程的成立不受内禀角动量的影响。从这个意义上讲,目前并没有充分的理由——哪怕只是理论上的理由——使人们必须在经典范围内拓展广义相对论的框架。

但是,将贝林番特张量引进爱因斯坦场方程的做法也并不是完全令人满意的。比如它使得表示角动量的能量动量起源的关系式 $J^{abc}=x^aT^{bc}-x^bT^{ac}$ 具有了完全的普遍性,而我们在前面提到过,量子意义下的自旋就不具有能量动量起源。因此,角动量与能量动量之间的这种关系式似乎不该具有那么大的普遍性,起码不该将量子意义下的自旋包括在内。而一旦认定量子意义下的自旋是一种与能量动量无关的角动量,那它对时空的影响就没有理由被包含在能量动量对时空的影响——即爱因斯坦场方程——之中。

另一方面,我们也不能简单地把自旋对时空的影响从理论中丢弃掉,因为虽然尚不存在自旋对时空产生影响的任何观测证据(考虑到自旋的微小,这是不足为奇的),但由于轨道角动量对时空的影响是广义相对论的确凿推论,在理论上单单把自旋对时空的影响丢弃掉无疑是极不自然的。这些都表明爱因斯坦-嘉当理论对自旋的处理——即既承认它对时空有影响,又不把这种影响归结于能量动量——是有一定合理性的。

除此之外,爱因斯坦-嘉当理论还有其他一些值得探讨的特点,比如它可

---

① 支持这种观点的一个重要理由是:从引力场的作用量原理所导出的场方程自动具有对称形式的能量动量张量。对这一点感兴趣的读者可参阅拙作《希尔伯特与广义相对论场方程》的第 3 节——收录于本书的"姊妹篇"《小楼与大师:科学殿堂的人和事》(清华大学出版社,2014 年)。

以将时空流形切空间上的结构群从广义相对论中的洛伦兹群(Lorentz group)推广到庞加莱群(Poincaré group)——这是嘉当提出这一理论的原始动机之一(我们所提及的量子意义下的自旋在当时尚未被发现),又比如它有可能对(部分地)消除广义相对论中的奇点问题起到一定帮助,等等。

不过,所有这些合理性及值得探讨的特点,都未能使爱因斯坦-嘉当理论得到太多的关注。原因在我看来有不止一条:比如爱因斯坦-嘉当与广义相对论的差别涉及到了像自旋这样的量子效应,从而不仅现在,哪怕将来也几乎没有任何可能得到直接的观测支持(引力在这种尺度上太过微弱)。此外,像有自旋粒子产生的引力场那样的问题,由于场源的量子特征无法忽略,很可能根本就不能用经典理论来处理①。假如经典理论根本就不能用,那么将广义相对论推广为爱因斯坦-嘉当理论的做法,也许就像当年索末菲(Arnold Sommerfeld)将玻尔理论推广为相对论性那样,缺乏真正的重要性。

2006年7月30日写于纽约

2014年12月13日最新修订

① 比方说,用广义相对论的克尔(Kerr)解来描述一个质量为 $m$,自旋为 $J$ 的微观粒子,将自旋视为角动量,则度规会在接近粒子康普顿波长(Compton wavelength)的 $J/m$ 处出现所谓的“裸奇环”。我们且不去理会那个很令人头疼的“裸”字——别想歪了,这是一个技术字眼,对之感兴趣的读者请参阅拙作《从奇点到虫洞》的第4章(清华大学出版社,2013年),在接近粒子的康普顿波长处出现像“奇环”那样的奇异性显然是不可接受的,也是与粒子物理实验完全矛盾的。虽然对微观粒子来说,我们原本就不该对经典描述有太多期待,但康普顿波长是经典与量子效应的分水岭,经典度规在“分水岭”上就出现如此巨大的问题,无疑是非常奇怪的,也是与引力在微观世界中的微弱性很不一致的。

# 黑洞略谈[①]

如果要在科学术语当中评选几个最吸引大众眼球的术语，黑洞（black hole）无疑会名列前茅。这个试图用引力把自己遮盖得严严实实的家伙不仅频繁出没于科幻故事中，而且在新闻媒体上也有不低的出镜率。前不久，一条有关美国国家航空航天局（The National Aeronautics and Space Administration，NASA）的“钱德拉”X 射线太空望远镜（Chandra X-ray Observatory）发现“最年轻黑洞”的新闻就被媒体竞相转载。而有关大型强子对撞机（Large Hadron Collider，LHC）有可能因产生微型黑洞而毁灭地球的传闻，更是不仅在过去几年时间里反复出现在各大媒体的显著位置上，而且还将美国和欧洲的司法界都卷入其中——因为有人试图通过法律手段来制止对撞机的启用，以“拯救”地球。在对撞机开始试运行的 2008 年 9 月，在印度甚至还发生了“一个‘黑洞’引发的血案”——一位 16 岁的花季女孩据说因担心微型黑洞毁灭世界而自杀。

---

① 本文曾以《有关黑洞的前世今生》为题发表于《中学生天地》2011 年 2 月刊（浙江教育报刊社出版）。

这个搅起了如此风波的黑洞究竟是什么东西呢？我们就围绕这两组新闻来谈谈它吧。

黑洞这个概念的起源通常被回溯到1783年，虽然那跟我们如今所说的黑洞其实没太大关系。那一年，英国地质学家米歇尔（John Michell）利用牛顿万有引力定律和光的微粒说推出了一个有趣的结果，那就是一个密度与太阳一样的星球如果直径比太阳大几百倍，它的表面逃逸速度将会超过光速。这意味着该星球对远方观测者来说将成为一颗“暗星”（dark star）——因为作为微粒的光将无法从它表面逃逸。不久之后（1796年），法国数学家拉普拉斯（Pierre-Simon Laplace）在其著作《世界体系》（*Exposition du Système du Monde*）中也提出了同样的结果。这个如今看来只有中学水平的结果，就是黑洞概念的萌芽。

但这个萌芽很快就枯萎了。

枯萎的原因是它所依赖的前提之一——光的微粒说在科学界失了宠，被所谓光的波动说所取代。光的波动说顾名思义，就是把光看成是一种波。但牛顿引力对这种波会有什么影响？却是一个谁也答不上来的问题。既然答不上这个问题，光能否从星球表面逃逸之类的问题也就无从谈起了。因此自《世界体系》的第3版开始，拉普拉斯悄悄删除了有关“暗星”的文字，他这个“与时俱进”的做法基本上为牛顿理论中的黑洞概念画上了句号。

黑洞概念的卷土重来是在20世纪的第二个十年。那时候，爱因斯坦（Albert Einstein）于1915年底提出了广义相对论（general relativity）。1916年初，一位被第一次世界大战的战火卷到前线，且罹患天疱疮（pemphigus），“阳寿”只剩五个多月的德国物理学家施瓦西（Karl Schwarzschild）得到了广义相对论的一个后来以他名字命名的著名的解——施瓦西解（Schwarzschild solution）。从这个解中，我们可以得到很多推论，比方说如果把太阳压缩成一个半径不到3千米的球体①，外部观测者就将再也无法看到阳光，这就是一种

① 更准确的说法是周长不到18.6千米（3千米×2π），因为那才是具有观测意义的量。但为行文方便起见，我们仍将使用“半径”这一术语，只不过它的真正含义是周长除以2π，而非径向距离。

现代意义下的黑洞——施瓦西黑洞。与米歇尔和拉普拉斯的“暗星”不同，现代意义下的黑洞具有很丰富的物理内涵，并且不依赖于像光的微粒说那样的前提①。

遗憾的是，施瓦西解的那些推论在很长的一段时间里不仅没有被人们所完全了解，反而遭来了一些针对黑洞的反对意见。就连爱因斯坦也曾提出过一些如今看来很幼稚的反对意见②。

不过“东边不亮西边亮”，另一个方向上的研究——即对白矮星（white dwarf）的研究——却殊途同归地将科学家们引向了黑洞。白矮星是耗尽了核聚变原料后的老年恒星，它们的质量与太阳相仿，块头却跟地球差不多，因而密度极高（一汤匙的白矮星物质的质量可达好几吨）。白矮星的发现给科学家们带来了一个问题：我们知道，恒星之所以能稳定地存在，是因为内部核聚变反应产生的巨大的辐射压强抗衡住了引力。但像白矮星那样不具有大规模核聚变反应的天体又是如何“维稳”的呢？这是一个很困难的问题。但幸运的是，当人们为这一问题伤脑筋时，一门新兴学科——量子力学——已经成熟了起来，在量子力学中有一条原理叫做泡利不相容原理（Pauli exclusion principle）。按照这条原理，电子是一群极有“个性”的家伙，每一个都坚持拥有独一无二的状态。如果你想压制这种“个性”，它们就会“殊死抗争”，这种抗争在宏观上会体现为一种巨大的压强，叫做“电子简并压”（electron degeneracy pressure）。白矮星主要就是依靠这种压强来抗衡引力的。当时

① 现代意义下的黑洞（施瓦西黑洞只是其中最简单的一种）与米歇尔和拉普拉斯的“暗星”很不相同，比如后者只是远方的观测者无法看到（由于作为微粒的光在“暗星”引力场中仍可运动一段距离，因此近处的观测者仍可看到），而前者则对于任何外部观测者都是“黑”的。

② 爱因斯坦计算了黑洞附近圆轨道上的粒子运动速度，结果发现轨道半径小于黑洞临界半径的1.5倍时，粒子运动速度会超过光速。他据此认为黑洞是不可能存在的。这一意见的幼稚之处在于，那计算无非说明在黑洞近旁粒子不可能维持圆轨道（除非有外力），而并不表示黑洞无法存在。这就好比在一个大漩涡里游泳者无法维持圆轨道，并不表示大漩涡不可能存在。

很多人认为，这就是恒星的终极“养老方案”，因为计算表明，“电子简并压”在任何情况下——即对于任何质量的恒星——都足以抗衡引力。

但一位印度年轻人无情地粉碎了这个美好的“养老方案”，此人名叫钱德拉塞卡（Subrahmanyan Chandrasekhar），本文开头提到的发现“最年轻黑洞”的“钱德拉”X 射线太空望远镜就是以他的名字命名的。

1930 年，本科刚毕业的钱德拉塞卡在研究白矮星时发现了一个出人意料的结果，那就是如果将相对论效应考虑在内，电子简并压将大为减弱，尤其是，当白矮星的质量超过太阳质量的 1.4 倍时，电子简并压将无法抗衡引力。可电子简并压是当时已知的能使老年恒星抗衡引力的唯一机制，如果这一机制不管用了，那老年恒星的命运会是什么呢？这一新问题使很多人深感不安，其中包括重量级的英国天文学家爱丁顿（Authur Eddington）。爱丁顿表示，钱德拉塞卡的结果是荒谬的，大自然是一定会让晚年恒星“老有所依”的。用今天的眼光来看，这是一种没什么说服力的单纯信念式的表态。不过在当年，这种表态却给钱德拉塞卡带来了很大的麻烦，他的论文直到一年多之后，才在遥远的美国找到一份杂志发表。

后来人们知道，恒星的“养老方案”其实不是唯一的，当电子简并压无法抗衡引力时，老年恒星还有另一种归宿，那就是中子星。这是一种密度比白矮星还高一亿倍（从而一汤匙物质的质量可达几亿吨）的天体，它依靠的是与电子简并压相类似、但更为强大的中子简并压。不过可惜的是，后者的强大也是有限度的，当中子星的质量超过太阳质量的 3 倍多时，中子简并压也会在巨大的引力面前败下阵来，这时的恒星就真的没救了，它的归宿只有一个，那就是黑洞[①]。因此黑洞不仅是施瓦西解（以及后来发现的若干其他解）的推论，更是大质量恒星演化的必然归宿。

---

① 有人提出过比中子星更致密的所谓“夸克星”（quark star）。不过“夸克星”即便存在，其密度也只会比中子星略大（如果说中子星像一个巨型原子核，那么夸克星就像一个巨型核子）。“夸克星”是否存在目前尚有争议，不过理论研究显示，无论它存在与否，都不太可能显著改变耗尽核聚变能量后大质量天体坍缩为黑洞的临界值。

但所有这些都只是理论，接下来的问题是：像黑洞那样“黑”的东西，如何才能得到观测上的证实？答案是：“解铃还需系铃人”，能帮助我们观测黑洞的，恰恰是那个使黑洞变“黑”的幕后推手——引力。黑洞虽然不发光，它的巨大引力却足以造成许多极为显著的观测效应，比方说，如果黑洞附近有足够多的物质，甚至有大质量的伴星，黑洞的巨大引力就会吞噬那些物质，而那些物质则会在掉进黑洞之前“垂死挣扎”——因剧烈碰撞等原因而发射出强烈的 X 射线（图 4）。探测这种 X 射线因此而成为了探测黑洞最重要的手段之一。

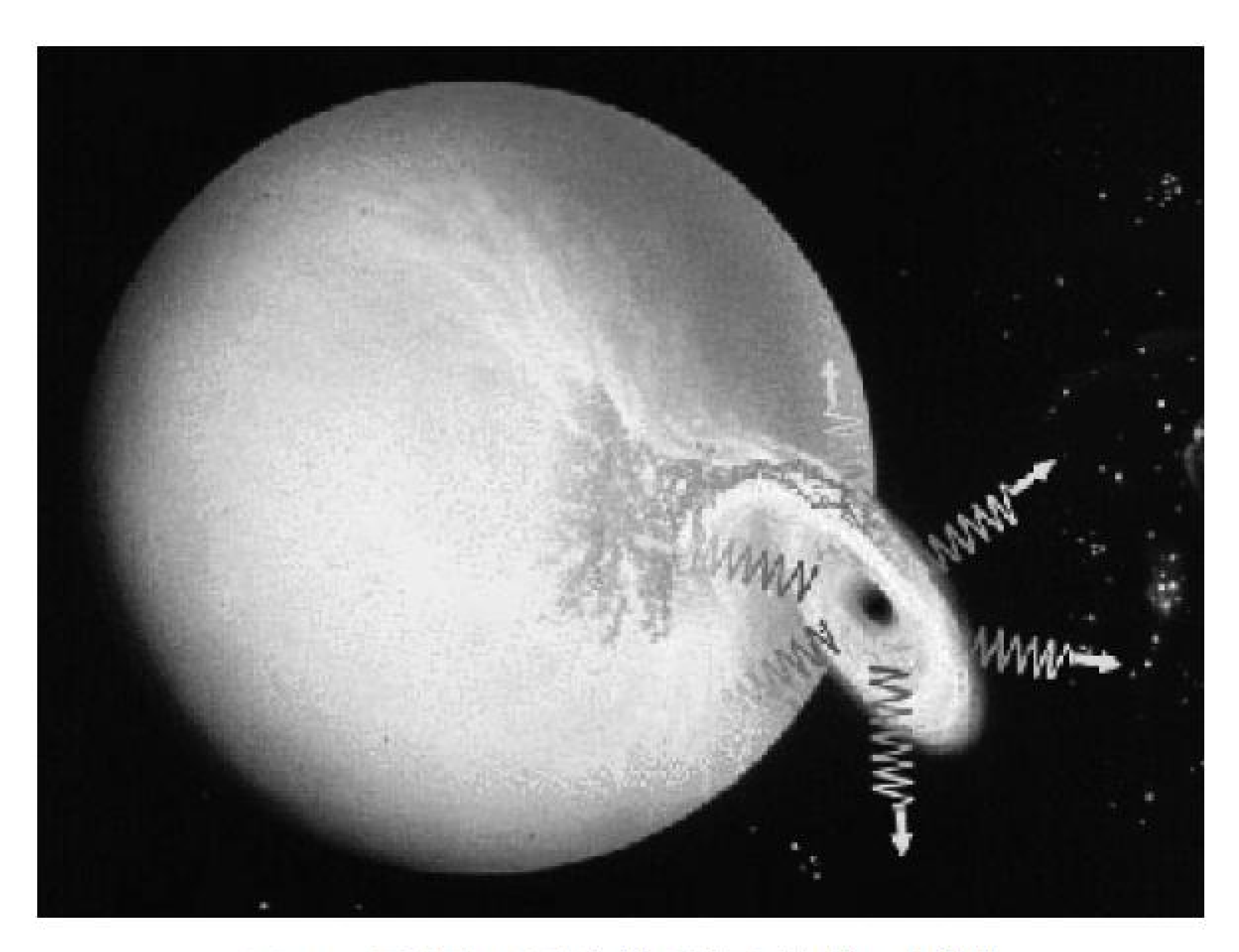


图 4　黑洞因吞噬物质而发射 X射线

好了，现在我们可以回过头来谈谈本文开头提到的那两组新闻了。“钱德拉”X 射线太空望远镜之所以能用来寻找黑洞，正是利用了物质在掉进黑洞之前会发射出强烈的 X 射线这一特点。而此次发现的黑洞之所以被称为“最年轻”，是因为它只有 30 多岁。我们怎么知道它只有 30 多岁呢？因为它是 1979 年观测到的一次超新星爆发的遗迹。不过要补充说明的是，这个黑洞位于距我们约 5000 万光年之遥的一个漩涡星系中，我们如今观测到的乃是它在 5000 万年前所发射的 X 射线，因此它的真正年龄其实是约 5000 万岁而不是 30 多岁。我们又怎么知道它是黑洞呢？那是因为天文学家们利用 X 射线能谱等数据估算了它的质量，结果约为太阳质量的 5～10 倍，超过了中子星的最

大可能质量①。这就是“最年轻黑洞”这一头衔的由来。

接下来再谈谈所谓大型强子对撞机有可能因产生微型黑洞而毁灭地球的传闻。大型强子对撞机是一个设计能量为 7 万亿电子伏特(7TeV)的对撞机(图 5)。那样的对撞机会产生黑洞吗？按照广义相对论，答案是否定的。因为这种万亿电子伏特(TeV)量级的能量在微观上虽然很高，用宏观标准来衡量却是微乎其微的，只不过是千万分之一焦耳的量级，这一丁点儿能量若想形成黑洞，除非是把它压缩到一个线度为一千亿亿亿亿亿亿分之一米($10^{-51}$米)的区域内，这比所谓的普朗克长度(Planck length)还小得多，与大型强子对撞机所能触及的最小线度相比，更是只有后者的一亿亿亿亿分之一($10^{-32}$)。因此按照广义相对论，大型强子对撞机是绝不可能产生微型黑洞的。



图 5　大型强子对撞机

既然如此，为什么仍有那么多人担心微型黑洞呢？因为他们背后有“军师”在指点，那些“军师”为他们的担心注入了一条重要理由，那就是在某些现代物理理论——比如超弦理论(superstring theory)——中，时空有不止四个

① 不过，由于对中子星最大可能质量的计算以及对“最年轻黑洞”的质量估算都有一定的误差，因此该天体究竟是黑洞还是中子星目前尚有一定的争议，只能说它有较大的可能性是黑洞。

维度。由于引力与时空密切相关,因此时空若有不止四个维度,引力的规律也将有所不同,而引力的规律一旦不同,产生黑洞的条件就会发生变化。理论计算表明,在那些带有额外维度的理论中,确实存在一些尚未被实验所排除的参数范围,使得大型强子对撞机有可能产生黑洞。

这么一来,事情就不太妙了。虽然那些认为时空有不止四个维度的理论目前还都只是假设性的,而那些使大型强子对撞机能产生黑洞的参数范围更是假设中的假设。但无可否认的是,有不少物理学家对那样的理论寄予厚望。因此,那样的理论所允许发生的事情即便只是假设性的,也不容忽视。毕竟,我们只有一个地球,实在不敢拿她去赌哪怕最细微的风险。

幸运的是,即便那些假设性的理论是正确的,并且参数也恰好处在能使大型强子对撞机产生黑洞的范围内,那样的黑洞依然是不可能毁灭地球的。因为黑洞还有一个我们尚未介绍的重要特点,那就是它并不是完全"黑"的。1974 年,英国物理学家霍金(Stephen Hawking)发现,由于量子效应的影响,黑洞会向外辐射能量。这种所谓的霍金辐射(Hawking radiation)对于大质量黑洞来说是微乎其微的,但对微型黑洞却极为显著,而且在时空有不止四个维度的情况下依然存在。计算表明,由于霍金辐射的存在,即便大型强子对撞机能够产生黑洞,那些黑洞也会在瞬息之间就"人间蒸发",别说毁灭地球,就连侵吞一两个原子都未必来得及。

至此,大型强子对撞机有可能因产生微型黑洞而毁灭地球的传闻似乎该烟消云散了。但事实却不然,有些人依然表示怀疑,因为霍金辐射尚未被观测证实过。虽然有关微型黑洞毁灭地球的担心本身也是建立在尚未被观测证实过的理论之上的,但当科学家们用同样类型的理论来回答他们的担心时,有些人却拒绝接受。对于这种近乎偏执的怀疑,有一样东西可以替科学家们作出回应,那就是宇宙射线。

大型强子对撞机是人类迄今所建能量最高的对撞机,但浩瀚的宇宙却有各种办法产生比那高得多的能量。观测表明,我们所栖居的地球每秒钟都会受到 10 万次以上的超高能宇宙射线的轰击,那些宇宙射线与地球物质发生碰

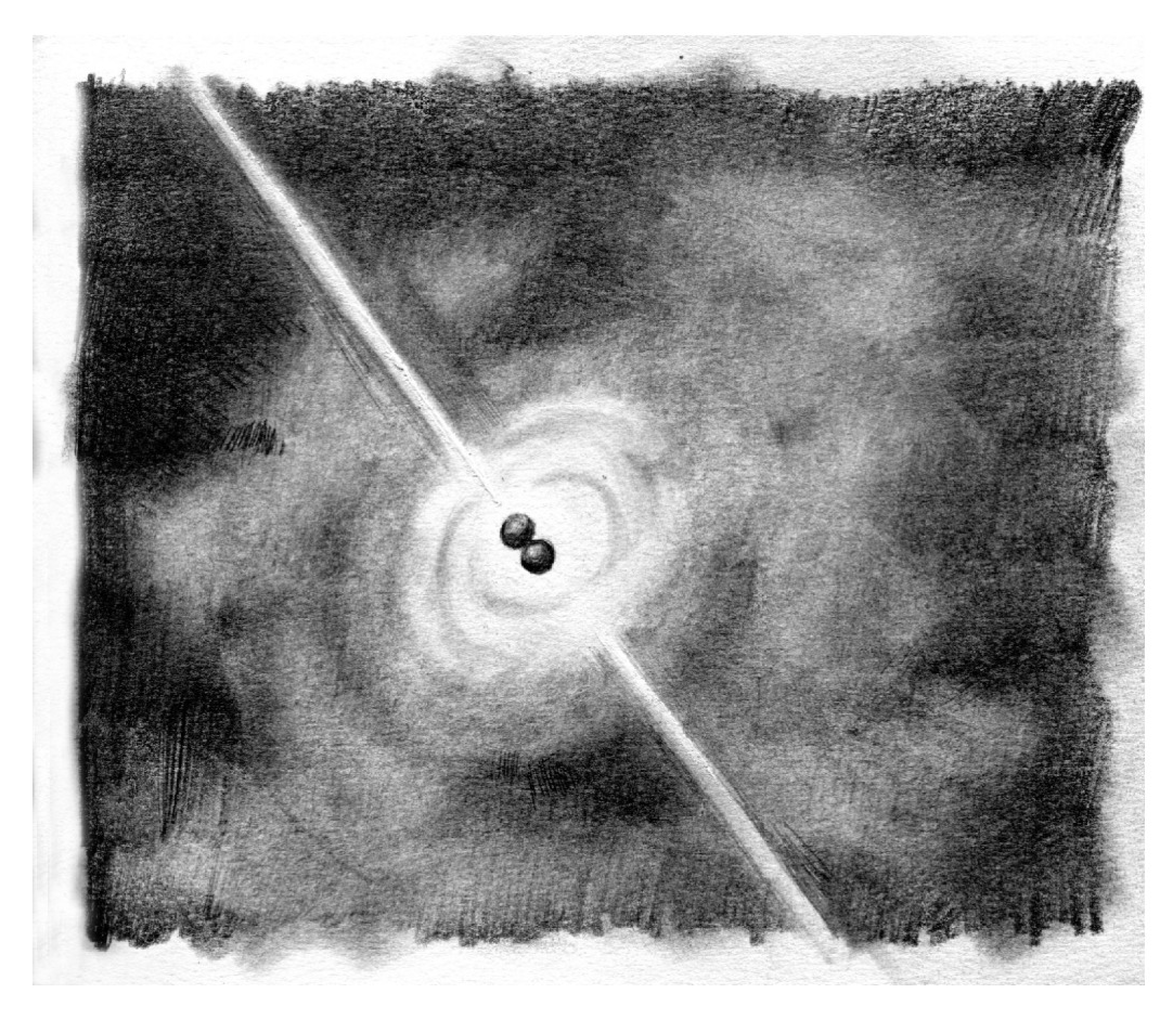
撞时所具有的能量比大型强子对撞机的能量更高①，而且那样的轰击自地球诞生以来，在长达45亿年的时间里从未间断过，相当于每时每刻都有大型强子对撞机在运行。如果大型强子对撞机果真有产生微型黑洞并毁灭地球的风险，无论其理论机制是什么，那样的风险都早该被宇宙射线转化为现实了。我们今天仍能坐在地球上争论这一问题本身，就很好地说明了那样的风险并不存在。事实上，如果我们把眼光放得更远一点，那么不仅地球每时每刻都受到大量超高能宇宙射线的轰击，表面积是地球一万多倍的太阳更是一个大得多的靶子，如果那样的轰击有危险的话，像太阳那样的庞然大物无疑会比地球死得更快。因此，包括太阳在内所有恒星的存在全都是极强的证据，表明大型强子对撞机因产生微型黑洞而毁灭地球的风险是完全可以排除的②。

事实上，大型强子对撞机若果真能产生微型黑洞的话，那不但不是什么风险，反而是了不起的实验成就，因为那不仅是对某些现代物理理论的绝佳检验，而且还是研究霍金辐射的最好、甚至有可能是唯一的直接手段。

2010年11月24日写于纽约

① 这个能量是指质心系能量。

② 严格讲，由高能宇宙射线产生的微型黑洞——如果有的话——与大型强子对撞机产生的微型黑洞有一个区别，那就是前者是高速运动的，从而会很快穿过地球。但研究表明，即便如此，假如那样的微型黑洞能够被产生，并且有毁灭星球的威力的话，宇宙中那些高度致密且具有强引力场的天体——比如白矮星和中子星——仍会因为俘获那样的黑洞而迅速灭亡，这同样与观测明显不符。



绘画：张京

# 反物质浅谈

## 一、一个令人苦恼的结果

众所周知,科幻小说作为一种特殊形式的小说,常从现代科学的发展中吸取新概念,反物质就是常被吸收的新概念之一。20 世纪 40 年代,美国科幻小说家威廉森(Jack Williamson)创作了一系列以反物质为题材的小说,称为 C. T. 故事,其中"C. T. "是他为反物质所拟的名称——"Contra-Terrene"——的缩写。威廉森的 C. T. 故事问世后不久,另一位美国科幻小说家阿西莫夫(Isaac Asimov)也在自己脍炙人口的机器人故事中引进了反物质的概念,他所设想的机器人大脑是所谓的"正电子脑"(positronic brain),而正电子乃是电子的反粒子,是反物质的基本组元之一。20 世纪 60 年代,著名科幻电视连续剧《星际迷航》(*Stat Trek*)开始播出,在这部连续创作和播出约 40 年之久、拥有不止一代忠实粉丝的电视连续剧中,反物质是星际飞船的重要燃料。这一点如今已几乎成为了所有以星际旅行为题材的科幻小说的共同特点。反物质概念在科幻小说中的频频出现,使公众对这一概念也产生了浓厚兴趣。

那么，反物质这一概念是何时，以何种方式被提出的？人们又是如何发现反物质的？反物质究竟是不是一种有效的星际飞船燃料？我们的宇宙中到底是物质多呢还是反物质多？这些或许是很多人不甚了解却不无兴趣的问题。本文将对这些问题作一些介绍。

反物质这一概念在学术界的出现最早可以追溯到19世纪末。1898年，英国物理学家舒斯特（Arthur Schuster）在给《自然》（*Nature*）杂志的一封信中提到，既然电荷可以有负的，金子说不定也可以有负的，而且负金子说不定和我们熟悉的金子有着一样的颜色。这或许是有关反物质的想法在科学文献中的萌芽。不过舒斯特有关反物质的想法只是一种简单而模糊的思辨，没有真正的理论依据，因而也没有引起任何重视。反物质概念在物理学上的真正渊源，是从将近30年后的1927年开始的。那一年，量子力学奠基人之一的英国物理学家狄拉克（Paul Dirac）提出了一个描述电子运动的数学方程。

狄拉克所提出的这一方程——即所谓的狄拉克方程（Dirac equation）——是一个既具有量子力学特征，又满足狭义相对论要求的方程，在当时是很令人耳目一新的结果①。更漂亮的是，这一方程还出人意料地自动包含了一些此前为解释实验结果而不得不人为添加到量子力学中的东西，一些在当时看来绝非显而易见的东西，比如电子的自旋和磁矩。作为一个方程式，狄拉克方程的形式之简洁，内涵之丰富，预言之神奇，似乎达到了物理学家们梦寐以求的境界。

但这一方程的“野心”似乎还不止于此，它还包含了另外一个重要结果——可惜这回却是一个令人苦恼的结果。

这个令人苦恼的结果是：狄拉克方程所描述的电子的总能量既可以是正的，也可以是负的。这个结果之所以令人苦恼，是因为人们在自然界中从未发

---

① 比狄拉克稍早，瑞典物理学家克莱因（Oskar Klein）、德国物理学家高登（Walter Gordon）及奥地利物理学家薛定谔（Erwin Schrödinger）也提出了一个试图融合量子力学与相对论要求的方程：克莱因-高登方程（Klein-Gordon equation）。但克莱因-高登方程具有一些当时看来比狄拉克方程更令人不易接受的特征，延后了它被真正重视的时间。

现过总能量为负的电子，因此狄拉克方程似乎允许存在一些自然界中不存在的东西。仅仅这样倒还罢了，因为允许存在的东西可以碰巧不存在，因此大不了假定自然界中所有电子的总能量碰巧都是正的。但不幸的是，按照量子力学，一个理论只要允许总能量为负的状态——即所谓的“负能量状态”，那么哪怕假定自然界中所有的电子的总能量碰巧都是正的，它们也会在很短的时间内通过量子跃迁进入到负能量状态，从而变成总能量为负的电子——也称为“负能量电子”。这种跃迁的结果无疑是灾难性的，与现实世界也大相径庭①。

## 二、错误描述中的正确结论

这么看来，狄拉克方程看似漂亮，实际上却似乎是错的，而且还错得相当离谱，足可把整个世界都搭进灾难里去。但是，狄拉克方程又分明包含了很多看起来正确得惊人的结果，一个错得如此离谱的方程又怎可能包含如此多正确得惊人的结果呢？莫非真的应了那句俗语：真理过头一步就是谬误？

为了解决这个令人苦恼的两难问题，狄拉克于 1930 年提出了一个大胆的假设，那就是负能量电子的确是存在的，不仅存在，而且还很多，多到足以把所有负能量状态都占满的地步。有人也许会问：既然有这么多负能量电子，为什么人们在自然界中从未发现过呢？答案是：由所有这些负能量电子组成的“海”就是我们平时所说的真空，从而不存在直接的观测效应。狄拉克之所以提出这样古怪的假设，是因为当时人们已经知道了一条重要的物理原理，叫做泡利不相容原理（Pauli exclusion principle），它表明任何两个电子都不能有相同的状态。既然任何两个电子都不能有相同的状态，那么一旦所有负能量状态都被负能量电子所占满，正能量电子也就不可能再通过量子跃迁进入到负

① 其实在经典相对论力学中也存在负能量状态，但在经典情况下我们可以摒弃负能量状态而不用担心它们对正能量状态产生影响，因为这两者之间存在一个非零的间隙（请读者想一想，对电子来说这一间隙有多大），而经典的物理过程都是连续的，从而不可能跨越这一间隙。

能量状态了。这样一来,负能量状态的存在也就不再成为问题了。

狄拉克的假设挽救了狄拉克方程,却带来了一个新问题。那就是他的假设虽然阻止了正能量电子进入负能量状态,却并不妨碍负能量电子因获得外来的能量而变成正能量电子。一旦出现这种情形,除产生一个正能量电子外,真空中还将出现一个因负能量电子空缺而形成的空穴,这种空穴等价于一个具有正能量,并且带正电荷的粒子。(请读者想一想这是为什么?)由此带来的新问题就是:这种带正电的粒子究竟是什么粒子呢?狄拉克的数学直觉告诉他那应该是一个质量与电子质量相同的粒子。但当时物理学家们所知道的唯一带正电的基本粒子是质子,其质量比电子质量大了 1 800 多倍。因此如果空穴所对应的带正电粒子的质量与电子质量相同,它将是一种新粒子,这是一个很大的麻烦。今天的读者也许难以理解这种视新粒子为麻烦的想法,因为换作是在今天,能够预言新粒子不仅不是麻烦,往往还会被认为是令人兴奋的结果(除非有显著的实验证据或理论依据表明所预言的新粒子不可能存在)。但提出新粒子这种后来一度成为家常便饭甚至蔚为时尚的做法,对当时的物理学家来说却几乎是一个思维禁区——一个连素以勇气著称的量子力学奠基者们也未敢轻易逾越的思维禁区。在这一思维禁区面前,具有极高数学天赋,并且一向崇尚数学美的狄拉克犯下了一生为数不多的显著错误之一,他放弃了自己的数学直觉,提出空穴对应的粒子是质子。

幸运的是,思维禁区束缚得了思维,却束缚不了计算;物理学家的思维禁区束缚得了物理学家,却束缚不了数学家。狄拉克的观点提出后,与他同时代的德国物理学家海森伯(Werner Heisenberg)和奥地利物理学家泡利(Wolfgang Pauli)分别对空穴的质量进行了计算,结果表明它应该与电子质量相同;德国数学家外尔(Hermann Weyl)更是从理论的对称性出发直接证明了这一点。另一方面,不管空穴是什么,既然它是电子离开所留下的,那么电子显然也可以重新跃回空穴,一旦出现这种情况,电子与空穴就会一起消失(变成能量),这种过程被称为湮灭(annilation)。如果空穴是质子,那么这就意味着电子可以与质子互相湮灭。这结果看起来显然很令人不安,因为电子

和质子是组成物质的基本粒子(当时中子尚未被发现),如果它们可以相互湮灭,那么物质的稳定性就成问题了。当然,问题到底有多严重还得看湮灭的快慢程度,或者说湮灭的几率。美国物理学家奥本海默(Robert Oppenheimer)和俄国物理学家塔姆(Igor Tamm)分别计算了这种几率,结果发现它相当大,足以使物质世界在很短的时间内就崩溃离析。

在这些结果的连环打击下,空穴是质子的假设遭到了灭顶之灾。1931年,狄拉克纠正了自己的错误,并提议将空穴所对应的质量与电子质量相同、电荷与电子电荷相反的实验上尚未发现的新粒子称为反电子(anti-electron)。这一回,他彻底突破了禁区,不仅提出了反电子,而且进一步提出质子及其他粒子——如果有的话——也应该有相应的反粒子。

如果所有的粒子都有反粒子,那么就完全有可能存在由反粒子组成的物质,这种物质就是人们所说的反物质。因此从某种意义上讲,这一年——即1931年——可以被视为是反物质概念诞生的年代。

按照狄拉克对反粒子的描述,反粒子是粒子脱离负能量状态后留下的空穴,因此反粒子与相应的粒子可以湮灭。这种湮灭有可能使粒子与反粒子同时转化为能量(比如光子)①,这是理论上所能达到的最高能量转化效率。这种转化效率是如此之高,以至于1克反物质与1克物质湮灭所产生的能量就足以超过“二战”末期美军投掷在日本广岛和长崎的两颗原子弹所释放能量的总和。不难设想,若有朝一日人类能广泛利用反物质作为能量来源,无疑将会带来巨大的技术飞跃。这是反物质成为很受科幻小说家们青睐的能量来源的根本原因。

不过需要指出的是,狄拉克对反粒子的描述虽然很直观,并且粗看起来颇有道理,在今天看来其实却只有历史价值,或者用美国物理学家施温格

---

① 正反粒子的湮灭产物可以是多种多样的。一般来说,参与湮灭的正反粒子的质量越大、能量越高,湮灭产物的种类通常就越多,在低能湮灭——尤其是轻粒子的低能湮灭——过程中,则有很大的几率产生光子对。

(Julian Schwinger)的话说,是“最好作为历史的猎奇而被遗忘”。为什么呢?因为如上文所介绍,狄拉克的描述需要通过泡利不相容原理来阻止正能量粒子进入负能量状态。对于电子和质子这样的粒子——被称为费米子(fermion)——来说,这恰好是可以做到的。但自然界中还存在另外一类粒子——被称为玻色子(boson),它们并不满足泡利不相容原理。对于那样的粒子,狄拉克有关反粒子的描述就无能为力了。不仅如此,按照狄拉克的描述,正反粒子的产生必须是成对的,因为一个新粒子的产生必定会留下相应的空穴——即它的反粒子;反过来说,新空穴的出现也只能是由于相应粒子的产生——即脱离负能量状态。但实验却表明这种粒子与相应反粒子的“双宿双飞”并不普遍成立。比方说在β衰变中,电子的出现就并不伴随有反电子。因此狄拉克对反粒子的描述细究起来并不正确,这一点不仅被多数科普读物所忽视,甚至在一些现代教科书中都没有明确说明,这是很有些不应该的。对反粒子的普遍描述,是在量子场论出现之后才建立起来的。不过狄拉克对反粒子的描述虽然并不正确,其所包含的一些基本结论,比如反粒子与相应的粒子质量相同,所带电荷及若干其他量子数相反,正反粒子可以相互湮灭,等等,却是普遍成立的,并且它的提出对量子场论的产生起到过启发作用,从这些意义上讲它对物理学的发展是功不可没的。

## 三、走错方向的电子还是走对方向的正电子?

与反粒子理论的曲折发展同样生动坎坷的,是实验物理学家们发现反粒子的故事。对于实验物理学家们来说,这个故事多少带着点遗憾,因为其实早在狄拉克提出反粒子概念之前,反粒子就已经在实验室里留下了踪迹,却被他们所忽略,这才让理论物理学家捷足先登。

在 20 世纪 30 年代,物理学家们探测带电粒子径迹的主要工具是云室(cloud chamber)。云室不仅可以显示带电粒子的径迹,通过将其置于磁场中,还可以进一步判断出粒子所带电荷的正负——因为正电荷与负电荷在穿

过磁场时会往不同方向偏转。早在狄拉克提出反粒子概念之前，实验物理学家们就在云室照片中发现过一些类似于电子，却与电子有着相反偏转方向的径迹。这些径迹其实正是反电子掠过云室留下的倩影。可惜就像狄拉克起初不敢把空穴诠释成反电子一样，实验物理学家们也未曾想到把那些反常径迹诠释成新粒子，从而错失了先于理论而发现反电子的机会。

直到狄拉克提出空穴是反电子之后，云室中那些反常径迹才引起了一些实验物理学家的重视。比如英国卡文迪许实验室（Cavendish Laboratory）的物理学家布莱克特（Patrick Blackett）就告诉狄拉克说，自己与同事可能已经发现了反电子存在的证据。但即便有狄拉克当出头鸟，布莱克特仍未敢贸然发表自己的发现，而是打算做进一步的核实。这一延缓将发现反电子的优先权拱手让给了大西洋彼岸的美国物理学家安德逊（Carl David Anderson）。

安德逊当时在美国西岸的加州理工大学（California Institute of Technology）从事宇宙射线研究。与其他一些实验物理学家一样，他也在自己的云室照片中发现了类似于电子，却与电子有着相反偏转方向的径迹，而且这样的径迹并不稀少，这一点引起了安德逊的重视，于是他把这一发现告诉了当时正在欧洲进行访问的导师密立根（Robert Andrews Millikan）。密立根是一位实验物理大师，曾因测量电子电荷及光电效应方面的工作获得 1923 年的诺贝尔物理学奖。对于安德逊所发现的径迹，密立根的解释是视之为质子产生的——质子所带电荷与电子相反，因而可以解释观测到的偏转方向与电子相反这一事实。但密立根的质子解释有一个致命的弱点，那就是像质子这样的重粒子在云室中的径迹应该远比像电子那样的轻粒子来得显著。可是安德逊所发现的径迹却并未显示出这种差异，因此密立根的质子解释很快被排除了。

另一方面，安德逊自己也提出了一种解释，他认为偏转方向与电子相反的径迹有可能是由反方向运动的电子产生的，这种解释也曾被欧洲物理学家们采用过。单纯从径迹的偏转方向上讲，它的确是能够说得通的。但安德逊的反向电子解释也有一个令人困惑的地方，那就是他所研究的是宇宙射线，而宇宙射线来自天空，从而应该是以大体相同的方向——即自上而下——穿越云

室的。既然如此,反方向运动的电子又从何而来呢?解决这一疑问最直接的办法无疑是对电子的运动方向进行直接检验。为此,安德逊在自己的云室中间插入了一片薄薄的铅板。由于粒子穿过铅板速度会变慢,因此只要对粒子在铅板上下的速度快慢进行比较,就可以判断出粒子的运动方向①。通过这一手段,安德逊发现绝大多数偏转方向与电子相反的粒子和电子一样来自天空,也就是说它们的运动方向与电子是相同而不是相反的。这就把安德逊自己的反向电子解释也排除了。

这两种解释都被排除了,留给安德逊的就只剩下一种解释了,那就是:他所发现的径迹来自一种带正电的、质量却远比质子轻的粒子——一种尚不被实验物理学家所知道的新粒子。但这种解释也有一个问题:那就是这样一个质量不大的新粒子为什么以前一直未被发现呢?如果安德逊知道狄拉克的空穴理论,他或许会想到那是因为这种粒子是反电子,它很容易因为与电子相互湮灭而从人们眼皮底下消失。可当时安德逊并不知道狄拉克的空穴理论,因此留给他的这唯一解释似乎看起来也不太可能。不过"看起来不太可能"和"不可能"终究是有差别的,福尔摩斯有一句虽不严谨但很管用的名言:当你排除了所有的不可能,剩下的无论看起来多么不可能,一定就是真相。安德逊知道这时候不应该犹豫了,于是他不顾密立根的反对,于 1932 年 9 月公布了自己的发现。

4 年后,这一发现为他赢得了诺贝尔物理学奖。

安德逊发现新粒子的消息一传到欧洲,布莱克特和他的同事立刻意识到自己犯下了迟疑不决的"兵家大忌",他们已经发现却未敢贸然发表的显然正是同样的粒子。于是他们立刻也发表了自己的结果。他们的结果虽不幸在时间上落后于安德逊,却有幸在空间上占据了一个有利条件,那就是他们离狄拉

① 在云室中比较同一种带电粒子的速度快慢是十分容易的,因为速度慢的粒子比速度快的粒子更容易被磁场所偏转,因此通过比较粒子径迹的偏转幅度——确切说是曲率——就可以比较出它们的速度快慢。

克很近。安德逊虽然发现了新粒子，却不知道它和电子的关系，而布莱克特和他的同事不仅知道新粒子和电子的关系，还知道它和电子可以成对产生，于是他们在自己的云室照片中有意识地寻找这种产生过程的证据，并如愿以偿地成为了首先发现正反粒子对产生过程的物理学家①。

在这些成果的发表过程中，反电子获得了一个新的、后来更为流行的名称：正电子（positron）。这个名称是一位杂志编辑向安德逊建议的，它的本意是“正子”（当时安德逊并不知道这一粒子与电子有关）。

## 四、从反粒子到反物质

正电子成为人类发现的第一种反粒子并非偶然。因为与之相比，其他反粒子要么在宇宙线及天然放射源中比较稀少，而早期加速器的能量又不足以产生；要么由于相互作用太弱而不易检测，其发现的难度都远远大于正电子。因此自正电子被发现之后，发现反粒子的步伐停顿了下来，直到二十几年后才迎来了一轮爆发。1955 年，意大利物理学家赛格雷（Emilio G. Segrè）与美国物理学家张伯伦（Owen Chamberlain）“领衔”发现了反质子（赛格雷和张伯伦获得了 1959 年的诺贝尔物理学奖）；次年，美国物理学家考克（Bruce Cork）及其合作者又发现了反中子。至此，组成物质的三种最重要粒子的反粒子都被发现了。此后，随着加速器能量的持续提高，其他基本粒子的反粒子也被陆续发现——当然，后来的那些发现对物理学家们来说已毫无悬念，因为在理论上，除少数粒子与自己的反粒子相同外，所有其他粒子都该有自己反粒子的观念早已确立。

不过尽管反粒子的发现和产生已不再稀罕，但反粒子很容易被“正”粒子

---

① 值得一提的是，当时和安德逊一同在加州理工大学跟随密立根从事实验物理研究的中国物理学家赵忠尧早在 1929 年至 1930 年间，就在研究硬 γ 射线穿越物质时，观测到了后来被证实为是源于正负电子对的产生的反常吸收效应，以及源于正负电子对的湮灭的特殊辐射——虽然这些实验并未直接观测正电子。

湮灭，因此如何保存它们依然是一个极大的技术难题。直到 20 世纪 80 年代，物理学家们才开始掌握了保存少量反粒子的手段。但是要想保存更多的反粒子，却又面临另一个技术难题，因为带同种电荷的反粒子相互排斥，中性的反粒子又不稳定。在这种情况下，要想积累反粒子，一种可能的手段是让反粒子像普通粒子配成原子那样配成中性的反原子。但是让那些极易湮灭，通常又高速运动的反粒子乖乖地组成原子又谈何容易？这项工作直到 1995 年才由德国物理学家欧勒特（Walter Oelert）领导的实验小组所完成，他们在欧洲核子中心（CERN）的低能反质子环（Low Energy Antiproton Ring）上成功地制备出了 9 个反氢原子。虽然只有区区 9 个，与普通原子动辄就是几个摩尔——1 摩尔约有 6 000 万亿亿（$6\times10^{23}$）个——的海量相比少得简直不值一提，但这一消息 1996 年初一经披露立即引起了世界性的轰动。许多大媒体用显著标题进行了报道，欧勒特本人也受到了媒体记者的“围追堵截”，有记者甚至试图把他从飞机上拦截下来进行采访。反氢原子的制备之所以引起媒体如此广泛的关注，一个很重要的原因是因为原子和分子是承载物质物理和化学性质的基本组元。从这个意义上讲，反氢原子的成功制备是人类有史以来首次制备出了反物质，此前所研究的只能称为是反粒子而不是反物质。对媒体来说，这无疑是一个极大的兴奋点。

不过欧勒特制备反氢原子虽是欧洲核子中心有史以来最受媒体关注的新闻之一，但该中心的粒子物理学家们却大都只是将之视为实验工艺上的成就，有人甚至戏称其为“新闻实验”。因为从理论上讲，由反粒子组成反原子乃是稀松平常之事；而从实用的角度讲，欧勒特制备的反氢原子不仅数量稀少，而且存在的时间也短得可怜，只有一亿分之四秒（$4\times10^{-8}$s），距离实用无疑还差得很远。欧勒特实验成功后的第二年，欧洲核子中心关闭了为这一实验及其他三十几个实验立下过汗马功劳的低能反质子环。这个低能反质子环在它服役的 14 年间总共产生了超过 100 万亿个反质子。如果把这些反质子全部当成反物质燃料与质子湮灭，它们所产生的能量大约可以让一盏 100 瓦的灯泡点亮 5 分钟。将这点微不足道的能量与 14 年间为产生这些反质子而消耗的

巨大能源相比，不难看到用反物质作为能源在目前还是极度得不偿失的。

但这些技术上的困难并不妨碍人类的想象力将反物质作为未来可能采用的一种能源。这种能源除了具有理论上最高的转化效率外，还有一个非常吸引人的优势，那就是洁净。我们知道，传统的能源，无论是化学能还是核能，通常都会在使用后产生有害的残留物，比如废气、核废料等，而正反物质的湮灭却可以将燃料彻底转化为能量，从而不留下任何残留物质，因此它是一种理论上最洁净的能源。这样既洁净又高效的能源不仅是科幻小说家的最爱，对于工程和军事领域来说也有着无穷的魅力。比如早在20世纪中叶，美国氢弹之父泰勒（Edward Teller）和苏联氢弹之父萨哈洛夫（Andrei Sakharov）就各自提出过反物质武器的可能性。在美苏冷战的后期，伴随"星球大战"计划的展开，美国军方开始了反物质应用方面的研究。

不过，反物质武器的制造除了有上面提到的困难外，还会面临一个意想不到的难题，那就是正反物质相互接触时，因湮灭而产生的辐射压会将正反物质剧烈推开，从而急剧减缓能量释放的速度。这种效应的一个"日常生活版"很多人也许早已见过，那就是：将一滴水滴在热锅上，水会渐渐蒸发，一般来说，锅越热，蒸发就越快，可是当锅热到一定程度后，水滴的蒸发状况会发生显著变化，它会在热锅上四处移动甚至跳跃，蒸发速度则反而大为减缓。这种有趣的现象早在两百五十多年前就被一位名叫雷登弗罗斯特（Johann Gottlob Leidenfrost）的德国医生注意到了，因而被称为雷登弗罗斯特效应（Leidenfrost effect）。雷登弗罗斯特效应的物理机制是：当锅热到一定程度后，水滴剧烈汽化产生的蒸汽会在水滴与锅之间产生一层蒸汽膜，阻隔两者的进一步接触，从而急剧减缓水滴的蒸发速度。这种机制也适用于正反物质的接触，只是蒸汽膜换成辐射层而已。雷登弗罗斯特效应对反物质武器的制造是一种障碍。不过，随着苏联的解体和冷战的落幕，近乎军事"大跃进"的反物质武器研究本就很快遭到了放弃。

到目前为止，除了基础物理研究外，反物质的主要应用领域是在医学影像方面。由于技术水平及反物质数量的稀少所限，多数其他类型的反物质应用

起码在目前还是很不现实的。不过，让想象力自由驰骋的话，未来的希望总是有的。比方说，假如宇宙中存在足够规模的天然反物质源，情况就将有所不同，因为那样我们就不必为制备反物质而费心了——虽然高效而安全地收集和保存反物质仍将是极具难度的挑战。

这就给科学家们提出了一个很大的问题，那就是：宇宙中有可能存在大规模的天然反物质源吗？

## 五、宇宙的主人和客人

物理学家们曾经对这一问题作出过肯定的猜测。狄拉克在他的诺贝尔演讲中就曾表示，如果正反物质是完全对称的，那么宇宙中完全有可能存在由反物质组成的星球。如果将这种猜测发挥一下，那么我们还可以设想宇宙中不仅存在由反物质组成的星球，甚至有可能存在由反物质组成的生物。另一方面，在宇宙大爆炸初期的极高温条件下，正反物质的产生应该是同等可能的，从这个角度讲似乎也有理由预期宇宙中存在大量的反物质，甚至在数量上与物质等量齐观。

但随着理论和观测的逐步深入，这些初看起来不无合理性的猜测渐渐冷了下来。

首先可以明确的一点是：由于反物质与物质会相互湮灭，因此在我们所生活的这颗小小的蓝色星球上，像发现煤矿或铀矿那样发现“反物质矿”是完全不可能的。不仅如此，反物质在整个太阳系中的存在也是微乎其微的，因为否则的话，由太阳发出，被称为太阳风的粒子流与反物质之间的湮灭早就应该被发现了。再往远处看，情况也没有实质的改变，虽然宇宙射线中存在一定数量的反粒子，有些地方甚至存在反粒子源，但那些反粒子大都来自普通物质所参与的高能物理过程。迄今为止并无任何确凿的证据，表明宇宙中可能存在反物质星球，或任何其他大范围的反物质分布。

事实上，不仅没有确凿证据表明宇宙中存在大范围的反物质分布，相反，

却有不少证据表明大范围的反物质分布不太可能存在。这种证据之一来自于宇宙中重子——主要是质子和中子——数量和光子数量的比值。我们知道，极早期宇宙中充斥着各种基本粒子，它们随时被高能物理过程所产生，也随时相互湮灭。当宇宙的温度逐渐降低时，粒子的产生过程开始受到抑制，因为它们所需的能量越来越难以达到。对于重子和反重子来说，这大致发生在宇宙温度为 10 万亿度的时候。在这个温度以下，湮灭过程起到主导作用，重子与反重子很快因为彼此湮灭而转变为光子或其他轻粒子。在那样的过程中重子与反重子变得越来越少，直至其密度低到连湮灭过程也无法有效进行为止，那时仍残留的重子就组成了我们今天所生活的物质世界（由此可见我们的物质世界是多么地来之不易）。这种过程所导致的一个显而易见的后果，就是今天宇宙中的重子数远远少于光子数，而且早期宇宙中的重子与反重子越对称，这种湮灭过程就会进行得越彻底，今天宇宙中的重子数相对于光子数也就会越少。观测表明，今天宇宙中的重子数与光子数的比值大约为 1 比 10 亿（$10^{-9}$）。这虽然已经是一个很小的比例，但理论计算表明，如果湮灭过程开始起主导作用时宇宙中的重子与反重子是完全对称的话，这个比值还要小得多，大约会是 1 比 100 亿亿（$10^{-18}$）。因此，我们所观测到的重子数与光子数的比值是一个很有力的证据，它表明早期宇宙中的重子与反重子是不对称的，而我们赖以生存的整个物质世界正是这种不对称的产物，是一个反物质极为稀少的宇宙。

有读者可能会问，是否有可能出现这样的情况，即早期宇宙中的重子与反重子完全对称，只不过由于某种原因而彼此分离了开来，从而没有发生有效的相互湮灭？如果是这样，那就既可以保持物质与反物质之间的对称性，又可以解释为什么我们观测到的重子数与光子数的比值远比由对称性所预期的 1 比 100 亿亿来得高。应该说，这是一个很不错的问题，事实上，物理学家们曾经考虑过这样的可能性。但这种猜测有两个致命的弱点：一是没有任何已知的物理过程可以将随机产生的重子和反重子有效地加以分离；二是如果早期宇宙中真的存在过这种正反物质分离的情况，那么正反物质的湮灭在空间分布

上将是高度非均匀的，这应该会在今天的宇宙微波背景辐射中留下遗迹。这样的遗迹并未被发现，因此这种可能性基本上可以被排除了。因此，无论观测还是理论都表明：我们今天所生活的宇宙是一个正反物质不对称的宇宙，物质是这个宇宙的主人，反物质只是稀客。

## 六、恼人的不对称之谜

既然我们所生活的宇宙是一个正反物质不对称的宇宙，那么一个很自然的问题就产生了，那就是为什么会出现这种不对称？对此，科学家们曾经有过两类不同的看法。其中第一类看法认为正反物质的不对称是由初始条件决定的，或者说是“先天”造就的。显然，这类看法比较消极，几乎等于是回避问题。令人欣慰的是，这种“偷懒”的看法在暴胀宇宙论出现后受到了沉重的打击。因为按照暴胀宇宙论，宇宙创生之初即便存在正反物质的不对称，也会在暴胀过程中被稀释得微乎其微。因此初始条件并不能对今天观测到的正反物质的不对称给出令人满意的解释。

既然初始条件不足以解释正反物质的不对称，那我们就只能寄希望于宇宙创生之后所发生的具体物理过程了，这就是第二类看法。这类看法认为我们今天观测到的正反物质的不对称是由某些特定类型的物理过程产生的。

那么，究竟什么样的物理过程才能造成正反物质的不对称呢？早在 1967 年，苏联氢弹之父萨哈洛夫就提出了那样的物理过程所需满足的三个条件：

（1）必须破坏费米子数守恒；

（2）必须破坏 C 和 CP 对称性；

（3）必须破坏热平衡。

这些条件后来被称为萨哈洛夫条件（Sakharov conditions），是任何能够产生正反物质不对称的物理过程或物理理论所必须满足的。

萨哈洛夫条件中的第一条提到的费米子是组成物质的基本粒子，比如电子、质子和中子（进一步细分的话，质子和中子是由夸克组成的，而夸克也是费

米子)。所有费米子的费米子数都是正的,而反费米子的费米子数则是负的。如果宇宙中的正反物质完全对称,那么总费米子数将是零。由于我们的宇宙中普通物质远比反物质多,因此总费米子数是正的。任何物理过程或物理理论要想让宇宙从正反物质完全对称(从而总费米子数为零)的状态演化到如今这个费米子数为正的状态,就必须改变总费米子数,从而必须破坏费米子数守恒。

萨哈洛夫条件中的第二条提到的 C 和 CP 对称性分别是基本粒子层次上的正反粒子对称性及正反粒子与宇称联合对称性。其中正反粒子对称性要求将一个物理过程中的所有粒子替换成相应的反粒子时,过程发生的几率不变。正反粒子与宇称联合对称性则是指在上述替换的同时再将物理过程换成它的镜像(好比是透过一面反射镜去看它)时,过程发生的几率也不变。这两个对称性之所以必须被破坏,是因为否则的话,任何可以造成物质多于反物质的物理过程都会伴随一个与它同样可能的、造成反物质多于物质的过程(即上述替换过程),这样两类过程的效果将会相互抵消。

最后,萨哈洛夫条件中的第三条之所以必须满足,是因为否则的话,任何可以造成物质多于反物质的物理过程都将与处在热平衡的逆过程相互抵消。

这三个条件虽被称为萨哈洛夫条件,不过萨哈洛夫本人在其长度只有三页的短文中其实并未如此鲜明地表述过这三个条件,这些条件是后人依据他的思路所归纳及重新表述的。

在这三个条件的基础上,物理学家们提出了许多理论模型,试图对正反物质不对称的起源作出定量解释。这些模型从相对简单的电弱统一理论(它是粒子物理标准模型的一部分),到各种各样的大统一理论,以及标准模型的超对称推广,种类繁多、应有尽有。但迄今为止,它们各自都存在一定的缺陷,或是结果的数量级不对,或是求解的困难度太大、或是过于特设、或是过于任意,尚无一个令人满意。不过尽管如此,现代物理为正反物质的不对称找到一个合理解释的前景看来是并不悲观的。

## 七、结语

我们有关反物质的介绍到这里就要结束了，虽然自人类发现反粒子迄今已有大半个世纪，但在理解物质与反物质的关系上还存在许多待解之谜。除了宇宙学尺度上正反物质的不对称外，在微观尺度上正反粒子也存在着令人困惑的不对称。物理学家们曾经认为，如果我们把一个微观物理过程中的所有粒子都替换成相应的反粒子，并且透过一面镜子去看它，那么我们所看到的新过程将与原过程有着相同的发生几率。这种对称性就是我们介绍萨哈洛夫条件时提到的 CP 对称性。由于这种对称性，反物质有时也被称为镜像物质。但令人困惑的是，这一对称性既非完全成立，也非完全不成立，而是非常接近成立①。大自然为什么要让这面特殊的镜子如此接近完美却又不让它真正完美呢？我们不知道。

反物质是宇宙中的稀客，但这稀客是从相对意义上讲的，宇宙中反物质的绝对数量依然是极其庞大的，足以为科幻小说留下巨大的驰骋空间，这是值得庆幸的。只不过，反物质星球的存在看来是极不可能的，因为没有任何天然的物理过程能够让反物质有效地汇集起来，并在这一过程中免遭普通物质的“致命骚扰”。而反物质生物的存在则比反物质星球更加不可能得多，因为即便存在反物质星球，在那种星球上要想演化出生物来也是难以想象的。我们知道，即便在距离太阳系的形成已有约 50 亿年、太阳系空间已相当“干净”的今天，地球每天仍会受到上千万次的陨石撞击（这些陨石绝大多数在大气层中烧毁，只有少数落到地上，因此我们不必担心它们会恰好砸在我们头上），这些陨石的总质量约有几吨。这样的质量相对于庞大的地球来说无疑是微乎其微的，

---

① 在 1957 年以前，物理学家们想当然地认为所有这类离散对称性都是严格的，直到 1957 年宇称对称性倒下之后，才开始对离散对称性进行区分，但它们大都像多米诺骨牌似地也倒下了。CP 是倒得比较慢的一个，前后也只经过了 7 年。

但同样的情形如果发生在一颗反物质星球上，那么这几吨的陨石（普通物质）与星球上的反物质湮灭所释放的能量将相当于上百万颗广岛原子弹爆炸所释放的能量①。要在一个每天被上百万颗原子弹轰击的星球上产生生物，这恐怕是最高级的想象力也难以胜任的。

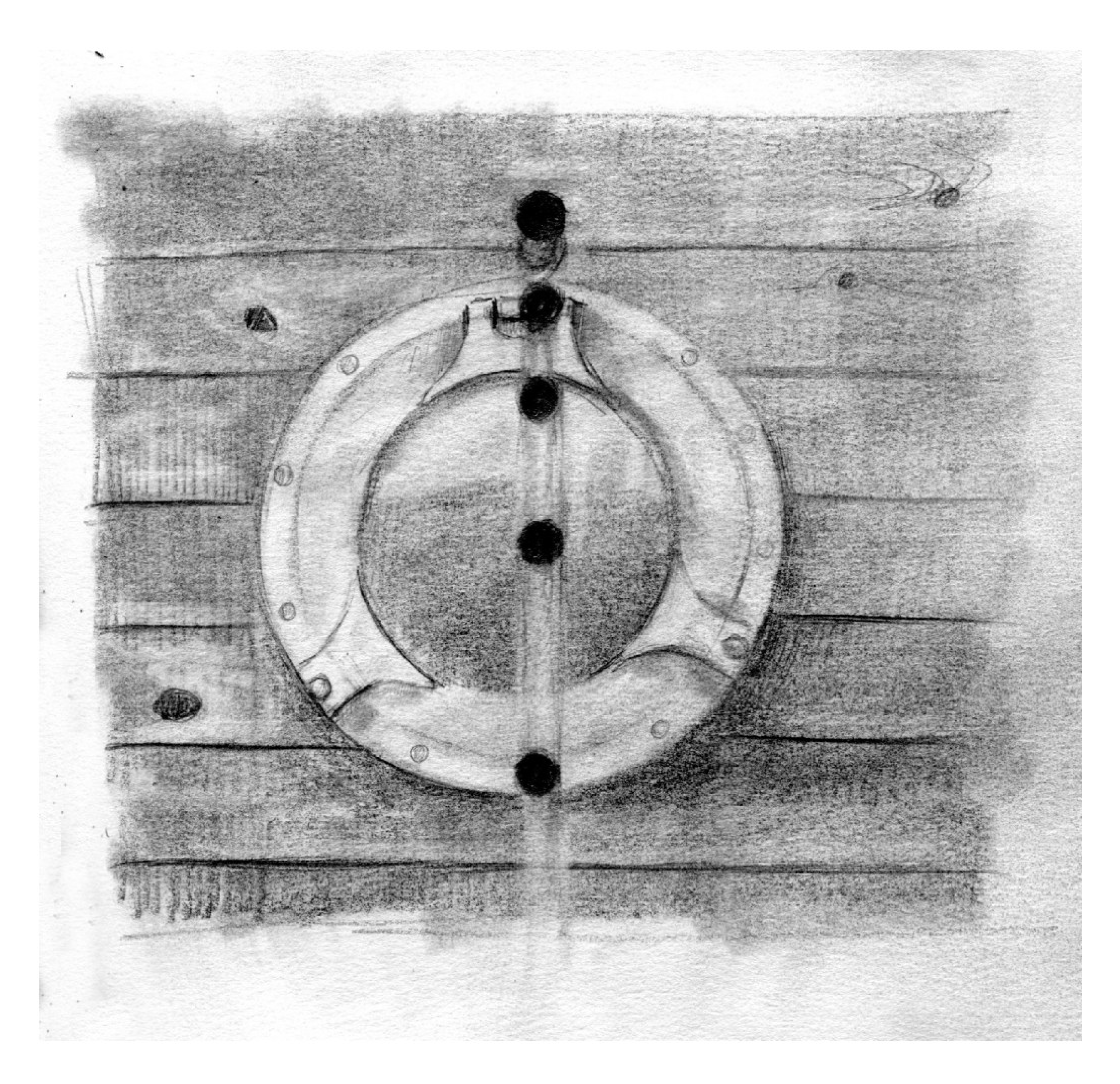
因此，如果有朝一日我们与某种外星球的高等生物建立了联系，我们可以大大方方地伸出手去和他们相握（如果握手对他们来说也代表友善的话），而不必担心大家会在这样的亲密接触中相互湮灭②。

2007 年 5 月 4 日写于纽约

2014 年 10 月 23 日最新修订

① 有读者可能会问：为什么不干脆假定那些陨石也是反物质？从纯粹假定的角度上讲，自然是可以的，但我们的讨论有一个前提，那就是承认我们这个宇宙——如目前的理论与观测所表明的——是一个物质为主的宇宙。在这样的宇宙中，越是大尺度的反物质分布就越不可能。因此我们对反物质出现的尺度只做最低限度的假定。

② 不过，如果我们真的担心他们有可能是反物质构成的，也有办法在见面之前加以确认，确认的方法就是利用刚刚提到过的微观世界正反粒子之间的不对称性。李政道在其教材 *Particle Physics and Introduction to Field Theory*（科学出版社出过中文版：《粒子物理和场论简引》）的第 9.2 节中对这一问题作了饶有趣味的论述，感兴趣的读者可以参阅。



绘画：张京

# 从伽利略船舱到光子马拉松[①]

## 一、从相对性原理到相对论

现代人都知道,我们脚下的大地并不是静止不动的。事实上,在读者们阅读本文标题的短短一秒钟的时间里,我们脚下的大地已随着地球的自转移动了几百米(除非你很靠近两极),随着地球绕太阳的公转移动了约 30 千米,随着太阳系绕银河系中心的公转移动了约 220 千米。而我们的银河系也没闲着,它相对于所谓的宇宙微波背景辐射参照系移动了约 550 千米[②]。这些运动大多数比火箭还快得多,人们却在很长的时间里一无所知,这是为什么呢?这个问题是我们的前辈在接受地球运动这一观念时面临的一大困扰,也是近代

---

① 本文是应《科学画报》约稿而写的关于破坏相对论的可能性的科普,原本有几段文字针对的是编辑指定的《新科学家》(*New Scientist*)杂志所报道的一个新理论,但由于该理论在所介绍的领域内并无特殊重要性,修订时我删去了与该理论有关的内容,使本文成为了一般性的介绍。

② 由于这些运动的方向各不相同,因此地球相对于宇宙微波背景辐射参照系的运动并不是上述数字的简单相加,而必须考虑方向的因素。观测表明,太阳系相对于宇宙微波背景辐射参照系的运动速度约为每秒 370 千米。(请读者想一想,我们为什么不给出地球的运动速度?)

科学的一个启蒙性的问题。

近代科学的先驱者之一，意大利物理学家伽利略(Galileo Galilei)在名著《关于两大世界体系的对话》(*Dialogue Concerning the Two Chief World Systems*)中对这一问题作了精彩的分析。伽利略注意到，地球运动的观念初看起来有违经验，其实却不然。相反，我们的经验表明，在一间封闭的船舱里，哪怕船在运动，只要运动得足够均匀，我们就无法发现它与处于静止时的任何区别。如果我们扔一块石头，往船头和船尾可以扔得一样远；如果我们观察一只小鸟的飞翔，它往哪个方向飞也都一样轻松。

我们现在知道，伽利略所注意到并归纳出的这一结果——即在所有匀速运动的参照系中，自然现象由相同的规律所支配——是一条非常重要的物理学原理：相对性原理(principle of relativity)。不过在伽利略之后两百多年的时间里，物理学的发展虽然迅速，相对性原理却不曾有机会展示它的真正威力。

但是到了19世纪末，情况有了变化。那时候，物理学家们遇到了一个恼人的问题，那就是当时最成熟的两类物理学规律——力学和电磁学规律——似乎不能同时满足相对性原理。或者换句话说，如果力学规律满足相对性原理，那么电磁学规律就不满足相对性原理，反过来也一样。这个“鱼和熊掌”的局面令人深感为难，考虑到力学规律满足相对性原理是自伽利略以来就被牢固确立的事情，物理学家们大都决定舍电磁学而取力学。但问题是：舍电磁学意味着电磁学规律不满足相对性原理，从而也就意味着我们能通过在伽利略船舱里做某些电磁学实验，来分辨轮船的运动。

情况果真如此吗？

还别说，物理学家们真的做了那样的实验，他们选择了一条很特殊的大船：地球。毫无疑问，这是一条运动的大船，这一点在19世纪末已是凡地球人都知道的常识了。物理学家们所做的实验是什么呢？是一个测定电磁波速度的实验。如果电磁学规律不满足相对性原理，那么电磁波沿不同方向的传播速度就会不一样——除非地球恰好是静止的。实验的结果是什么呢？让人

大跌眼镜，地球竟然真的是静止的！这下麻烦大了，难道兜了几个世纪的大圈子，我们又要重回地心说的年代？

幸运的是，这时有位名叫爱因斯坦（Albert Einstein）的专利局职员及时作出了一个相反的选择：舍力学而取电磁学。这样一来，所有证明地球静止的电磁学实验就都不再有效，比方说测定电磁波速度的实验就会像在伽利略船舱中扔石头一样的无效。而我们——谢天谢地——也就不必重回地心说的年代了。但问题是：既然舍了力学，那力学规律该怎么办？爱因斯坦的回答很简单，那就是“削足适履”。既然力学规律这只脚放不进与电磁学规律相一致的相对性原理那只鞋，那就修改力学规律。

修改力学规律的结果是导致了一些很新奇的结果，比方说物体的质量原本被认为是常数，修改之后却变成与相对运动有关的了。爱因斯坦的这一回答实际上是把相对性原理提升为了一条比像力学、电磁学那样具体领域的物理理论都更基本的原理，由此建立的理论就是所谓的相对论（theory of relativity）。相对论在更广阔的背景下再次确立了伽利略的观察，即在伽利略船舱中所做的任何实验或观测，都不可能分辨轮船的运动。

在此后一个多世纪的时间里，得到无数实验验证的相对论成为了现代物理学最坚实的基石之一。我们描述基本粒子的理论被称为相对论量子场论（relativistic quantum field theory），我们描述宇宙的理论被称为广义相对论（general theory of relativity）①，我们描述日常现象的力学、电磁学等也全都满足相对论的要求。而当年的专利局职员则成为了有史以来最伟大的科学家之一。

一切似已尘埃落定。

但是，物理学家们注定是一群不安分守己的人，新的探索无论对于他们的

① 在广义相对论的每个时空点附近足够小的区域内，都可以找到特殊的参照系，在其中物理规律与在匀速运动的参照系中一样，这就好比光滑曲面上每个点附近足够小的区域都很接近平面一样。

好奇心还是职业都是必不可少的。相对论无疑是一座巍峨的高山，但物理学家们仍然要问：山的那边还有没有风景？

## 二、破坏相对论的思路与后果

物理学家们之所以要这样问，当然也有具体的原因。比方说我们前面提到的两个理论——描述基本粒子的相对论量子场论与描述宇宙的广义相对论——虽然各自都很成功，却迄今无法和睦共处。更糟糕的是，作为物理理论，它们又不可能做到井水不犯河水。因为在有些场合——比如在大质量、高密度的天体附近——哪怕是基本粒子之间的相互作用，也必须考虑引力的影响；又比如在宇宙大爆炸的初期，整个宇宙都处在微观尺度上，哪怕是最宏观的性质，也不能忽略量子效应。因此，相对论量子场论与广义相对论必须以某种方式融合到一起，这种融合是现代物理学所面临的最棘手的课题之一。

有意思的是，试图将这两个同时满足相对论要求的理论融合到一起的努力，却为破坏相对论的可能性开启了思路。

其中有一种努力的途径是认为问题的根源在于时空貌似光滑，其实却不然。当我们探索到只有原子核的一万亿亿分之一（$10^{-20}$）的尺度——被称为普朗克尺度——上时，时空也许会显示出像网格一样的结构。这就好比一片丝绸，远远看去很光滑，拿到放大镜下，却可以看到密密层层的网格结构。如果时空真的有那样的网格结构，那么伽利略船舱中的人只要有足够厉害的“放大镜”，就有可能通过观测时空的网格结构，来判断轮船是否在运动，从而破坏相对论的要求。

另一种努力的途径则是认为，时空中有可能存在一种被称为“背景场”的东西。这种东西不是由物质产生的，却能对物质施加影响（用物理学家们的术语来说，这是一种非动力学场），而且这种影响在不同位置、不同时刻，甚至对不同观测者都有可能是不一样的。如果说时空网格像一片丝绸，那么这种背景场就像一种流体——比如水。在水中，即便我们无法像观察丝绸网格那样

观察水分子，也依然可以判断物体的运动，因为我们可以观察水对物体的阻力。如果时空中真的存在那样的背景场，那么伽利略船舱中的人就可以通过观察它对普通物体的作用来判断轮船是否在运动，这同样破坏相对论的要求。

上面这些思路并非单纯的幻想，而是多少有一些物理上的缘由，甚至是某些理论模型的推论。比如时空的网格结构与一种被称为“圈量子引力”(loop quantum gravity)的理论不无渊源，而背景场的思路则可以从所谓的“超弦理论”(superstring theory)中获得某种支持①。

破坏相对论这个潘多拉盒子一经打开，其他可能性也就应运而生了。比如有一种思路是这样的：将现实世界的物质全都扔掉，直接对相对论的数学结构开刀，由此可以得到一种被称为“双重狭义相对论”(doubly special relativity，DSR)的理论。这是一种很大胆的思路，可惜的是，迄今还没人知道如何将被扔掉的物质重新放回到理论中去，因此这种思路的物理意义起码在目前还是成问题的②。不过在一个连相对论都被怀疑的研究方向上，谁又敢说这种思路一定就没有可能呢？历史上纯粹源自数学考虑，却最终获得物理意义的例子毕竟还是有的，因此这样的思路也有一些人在研究。

看来破坏相对论的思路不仅有，而且还不止一条。

既然如此，那就让我们姑且假定相对论果真被破坏了。接下来的一个很重要的问题是：这种破坏会有什么后果？对这个问题的具体答案显然跟破坏

① 超弦理论本身是符合相对论要求的——确切地说是具有洛伦兹对称性(Lorentz symmetry)的，超弦理论中的相对论破坏(确切地说是指破坏洛伦兹对称性)是以对称性自发破缺的形式出现的。

② 具体地说，双重狭义相对论是通过对动量空间中的庞加莱代数(Poincaré algebra)进行修改而来的，因此有时也被称为变形狭义相对论(deformed special relativity，缩写恰好仍是DSR)。双重狭义相对论除了像狭义相对论一样存在一个不变速度外，还存在一个不变动量(名称中的“双重”一词便由此而来)。双重狭义相对论的部分特点可以在某些非对易几何模型中找到渊源(但也只是数学渊源)，另有些人则希望(目前还只是奢望)它能与圈量子引力建立联系。但迄今为止，该理论只有运动学，而无动力学，甚至连自洽性都尚待澄清。

相对论的具体方式有关，不过，由于破坏相对论的思路大都与时空的结构有关，而时空是引力的源泉，因此我们可以预期，破坏相对论的后果之一，就是使引力发生变化。

比方说，如果破坏相对论的肇事者是背景场，就有可能对引力产生影响。我们在前面提到过，背景场能对物质施加影响，这种影响的可能的体现方式之一就是对引力的修正。而且这种修正在不同位置、不同时刻可以是不同的——或者用一些科普报道所用的比喻来说，是苹果在不同季节的掉落快慢有可能是不同的。

除了苹果的掉落快慢有可能不同这样的"家常"后果外，破坏相对论还可能造成一些更严重的后果。比方说，相对论中有一条很基本的原理，叫做光速不变原理①，它表明光速是一个普适的极限速度。在很多破坏相对论的理论中，这条原理不再成立，不同的粒子可以有不同的极限速度。初看起来，这似乎没什么大不了，但是有科学家研究后发现，利用这一结果可以在黑洞附近让热量自发地从低温物体传向高温物体②。这是一个令人吃惊的结果，因为在自然界中，热量的自发传输一向是从高温物体传向低温物体，而不能相反。这是一条很重要的物理学原理，叫做热力学第二定律，违反这一原理的物理过程被称为第二类永动机，它与违反能量守恒定律的第一类永动机一样，被认为是不可能实现的。

因此，破坏相对论的后果很可能是牵一发动全身的，它所引发多米诺骨牌

① 这条原理是让电磁学规律满足相对性原理的必然推论。

② 这是2006年俄罗斯科学院核子研究所(Institute for Nuclear Research of the Russian Academy of Sciences)的两位物理学家在《物理快报》(*Physics Letters*)上发表的一个结果。他们的大致思路是这样的：不同的粒子具有不同的速度上限意味着黑洞辐射中不同的粒子会有不同的辐射温度。假定粒子B的辐射温度高于粒子A，我们在黑洞外面构筑两个壳层，壳层A只能发射和吸收粒子A，壳层B只能发射和吸收粒子B，我们选择壳层的温度使得(粒子B的辐射温度)>(壳层B的温度)>(壳层A的温度)>(粒子A的辐射温度)。在这样的安排下，壳层A会通过粒子A将热量传给黑洞，而黑洞又会通过粒子B将热量传给壳层B，净效果是壳层A将热量传给壳层B，即热量自发地从低温物体传往了高温物体。

效应，很可能导致其他一些很重要的物理学原理也被破坏。这其实是可以预期的，因为物理学是一个整体，它的各个分支之间有着千丝万缕的关联，它的基础并不是一系列孤立假设的集合，我们很难在破坏像相对论那样的重要部分时不影响到其他部分。

## 三、光子的马拉松——破坏相对论的证据？

以上我们介绍了很多理论上的东西，在物理学上，再雄辩的理论也离不开观测与实验的评判。对于相对论的破坏来说，它即便存在也极其微弱，我们该如何去寻找观测与实验的评判呢？在当前的条件下，比较有希望的探索方向主要有两类。

一类是探索微观世界的对称性破缺。这类探索有一段不短的历史。在1957年以前，人们曾经以为微观世界充满了对称性，其中很重要的一条是说微观世界的规律可以通过一面镜子去看而不被改变——这被称为宇称(parity)对称性。可惜这一对称性在1957年被证实是破缺的——确切地说是在所谓弱相互作用中是破缺的。不过这一对称性还可以加强，比如在通过镜子去看的同时把粒子与反粒子对换，可惜就连这种加强版的对称性在1964也被证实是破缺的——也是在所谓弱相互作用中破缺。但这一对称性还有一个终极加强版，那就是在通过镜子去看的同时，不仅把粒子与反粒子对换，而且让时间倒流。一些理论研究表明，在某些合理的条件下，这种终极加强版的对称性与相对论几乎是“一条绳上的两只蚂蚱”，一旦前者遭到破坏，后者也难以独善其身①。按照这一结果，只要我们能在微观世界里找到任何确凿的现象破坏这种终极加强版的对称性——比如发现任何一个基本粒子的质量、自旋、电荷、衰变方式等性质与反粒子不严格对应——就相当于间

① 但反过来则不然，即相对论的破坏不一定意味着那种终极加强版的对称性——即所谓的CPT对称性——的破坏。因此严格地讲，它们并不完全是“一条绳上的两只蚂蚱”。

接证实了相对论的破坏。这方面的实验数据可以说是天天都在积累(虽然目的大都不是为了证实相对论的破坏),但迄今尚无任何证据显示相对论被破坏。

另一类探索在思路上更为直接。我们刚才提到过,在很多破坏相对论的理论中,光速不变原理不再成立。由此导致的结果,是不同的粒子可以有不同的极限速度。但除此之外,它往往还意味着不同能量的光子在真空中的传播速度彼此不同——这被称为真空色散(vacuum dispersion)。利用这一特点,我们可以让不同能量的光子进行跑步比赛,来观察它们的速度是否不同,进而判断相对论是否被破坏①。不过由于光子的速度实在太快,彼此的速度差异又即便有也极其细微,要想分出胜负,比赛必须是马拉松,而赛场只能是星空。

2005 年夏天,天文学家们终于观察到了这样一次马拉松,一群高能光子从一个编号为“马卡良 501”(Markarian 501)的遥远的活动星系核出发,经过 5 亿年的漫长旅程,抵达了地球。这群光子是一次伽马射线耀斑(gamma ray flare)的产物,它们的抵达被位于西班牙西南加那利群岛(Canary islands)上的“大气伽马切伦科夫成像望远镜”(major atmospheric gamma-ray imaging Cherenkov telescope,MAGIC)所记录。在记录中令科学界感到震动的是,能量在 1.2～10TeV 之间的高能光子的到达时间比能量在 0.25～0.6TeV 之间的低能光子晚了约 4 分钟,这与某些破坏相对论的理论所预期的大致相符。

那么,我们是不是可以就此宣布相对论被破坏了呢?不能。因为我们对这场 5 亿年前就起跑的马拉松知道得还太少,高能光子的到达时间虽然晚了 4 分钟,但它的起跑是否也晚了呢?我们却一无所知。

---

① 确切地讲,不同能量的光子具有相同速度可以推翻许多破坏相对论的理论,但相反的结果,即不同能量的光子具有不同速度,却并不能直接证实相对论的破坏,因为相对论所要求的只是存在一个不变速度,这个速度不一定非得是光子的速度,甚至不一定非得有任何粒子具有这一不变速度。

而更有意思的是，2009 年，科学家们通过翱翔在外层空间的“费米伽马射线太空望远镜”（Fermi gamma-ray space telescope，FGST）又观测到了一次光子马拉松（图 6）。参加这次马拉松的光子来自一次伽马射线暴（gamma ray burst），它的威力比产生前一次马拉松的伽马射线耀斑还要巨大得多，距离也更遥远得多（红移值约为 0.9）。那些光子经过了数量级为百亿年的漫长跋涉才抵达地球，这几乎是我们这个宇宙所能提供的最长的赛程。这赛程是如此之长，以至于在这次马拉松起跑的时候，不仅我们不存在，就连我们脚下这颗蓝色星球都尚未形成！与上次不同的是，这次马拉松的结果是高能光子（能量约为 31GeV）与低能光子（能量在 10keV 以下）几乎同时到达终点（时间差在几十毫秒到几秒之间，几乎可以忽略），从而不仅没有破坏相对论，反而几乎给所有破坏相对论的理论下达了死亡通知书①。



图 6　费米伽马射线太空望远镜

两次光子马拉松，一对彼此相反的结果，我们究竟该相信什么呢？答案恐怕是：什么都先别相信，去寻找更多的证据。著名的美国行星天文学家萨根（Carl Sagan）有一句名言：超常的主张需要超常的证据（extraordinary claims require extraordinary evidence）②。在相对论所具有的庞大的证据链面前，破

① 因为如果这次光子马拉松的结果可信，那么破坏相对论的效应将会细微到不自然的程度，比方说对于最简单的真空色散模型——即色散率的修正项线性正比于能量的模型——来说，破坏相对论的能标将会比所谓的“普朗克能标”（Planck scale）还高得多。

② 萨根的这一表述具有较大的公众影响，不过他并不是最早提出这类原则的人，早在两百多年前，法国数学家拉普拉斯（Pierre-Simon Laplace）就曾说过：“支持一个超常主张的证据分量必须正比于主张的奇异程度。”

坏相对论的理论无疑是超常的主张，但那两次光子马拉松却绝非超常的证据（更不用说它们还彼此矛盾），对所有有志于这一领域的研究者来说，探索的路还很漫长。

2009 年 9 月 25 日写于纽约

2014 年 11 月 9 日最新修订

# 质量的起源[①]

## 一、引言

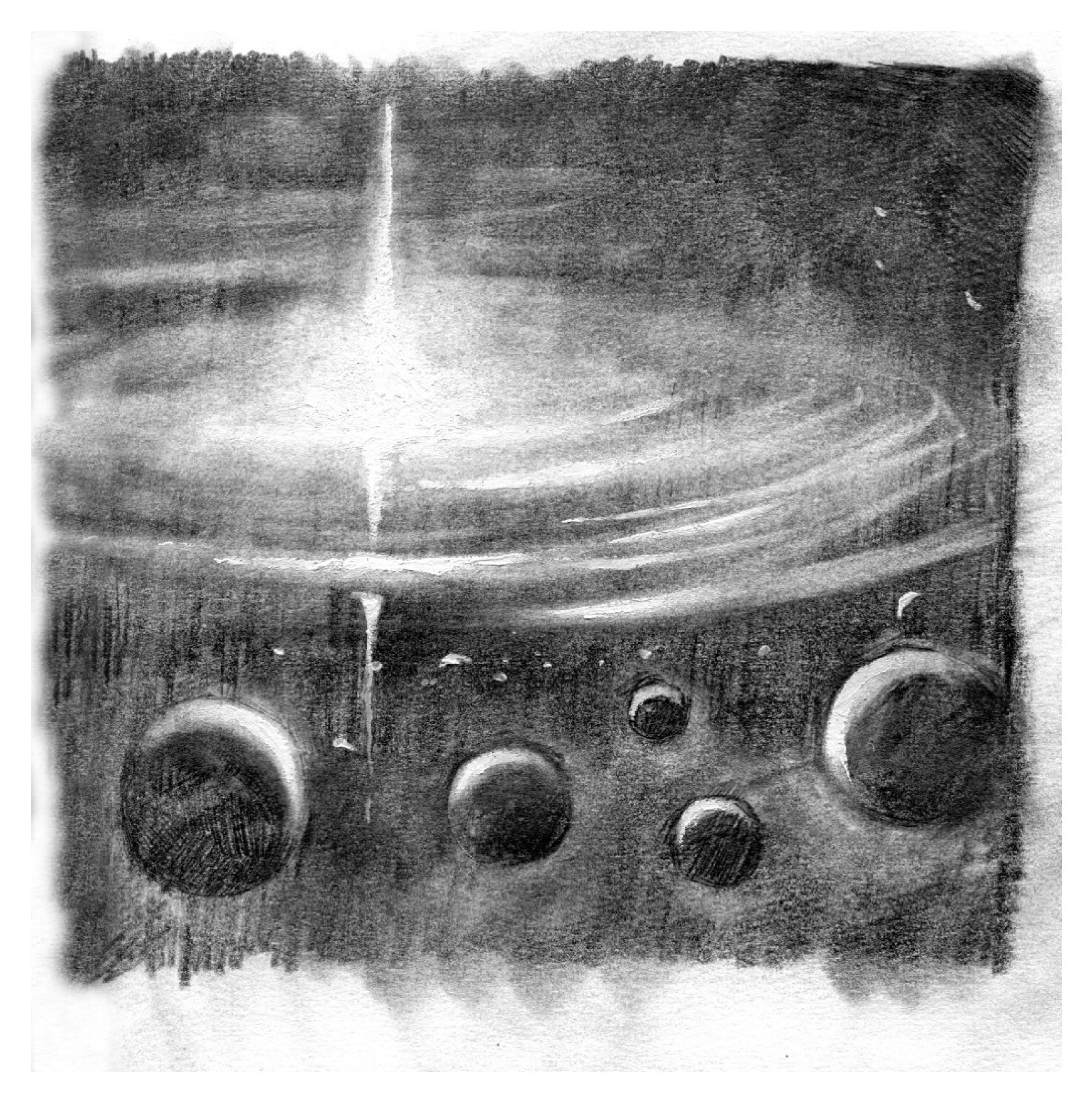
物理学是一门试图在最基本的层次上理解自然的古老科学，它的早期曾经是哲学的一部分。在那个时期，物理学所关心的是一些有关世界本原的问题。那些问题看似朴素，却极为困难。在后来的漫长岁月里，物理学曾经一次次地回到那些问题上来，就像远行的水手一次次地回望灯塔。

“质量的起源”便是一个有关世界本原的问题。

## 二、宇宙物质的组成

我们首先来界定一下所要讨论的质量究竟是什么东西的质量。这在以前是不言而喻的，现在的情况却有了变化，因此有必要加以界定。众所周知，过

① 本系列曾在《现代物理知识》杂志（中国科学院高能物理研究所）上连载，其中第3～6节发表于2007年第1期（发表时的标题为：《质量起源——电磁质量说的兴衰》）；第7～8节发表于2007年第2期（发表时的标题为：《质量起源——从对称性破缺到希格斯机制》）；第9～13节发表于2007年第3期（发表时的标题为：《质量起源——量子色动力学与质量起源》）。



绘画：张京

去十年里观测宇宙学所取得的一个令人瞩目的成就，就是以较高的精度测定了宇宙物质的组成，从而使我们在宇宙学的历史上第一次可以谈论所谓的“精密宇宙学”（precision cosmology）。

按照这种“精密宇宙学”为我们绘出的图景，在宇宙目前的能量密度中暗能量（dark energy）约占 68%，暗物质（dark matter）约占 27%，而我们熟悉的所谓“可见物质”（visible matter）或“普通物质”（ordinary matter）只占可怜兮兮的 5%。在这些组成部分中，对暗能量与暗物质的研究目前还处于很初级的阶段，尚未建立起足够具体且有实验基础的理论。因此本文对之不做讨论。

除去了暗能量与暗物质，剩下的就是可见物质了。可见物质在宇宙能量密度中所占的比例虽小，却是我们所熟知的物质世界的主体。可观测宇宙中数以千亿计的星系，每个星系中数以千亿计的恒星，以及某个不起眼的恒星附近第三颗行星上数十亿的灵长类生物，全都包含在了这小小的 5%的可见物质之中①。

本文要讨论的便是这可见物质。

与“暗”字打头的其余 95%的能量密度相比，我们对可见物质的研究与了解无疑要深入得多。今天几乎每一位中学生都知道，这部分物质主要是由质子、中子、电子等粒子组成的。因此很明显，要讨论质量的起源，归根到底是要讨论这些粒子的质量起源。

## 三、从机械观到电磁观

对几乎所有受过现代教育的人来说，最早接触质量这一物理概念都是在牛顿力学中。在牛顿力学中，质量是决定物体惯性和引力的基本物理量，是一个不可约（irreducible）的概念。我们知道，在大约两百年的时间里，牛顿力学

① 当然，这一说法并不严格，在星系所占据的空间范围内也有数量可观的暗物质及暗能量，我们这里指的只是光学观测意义上的星系。

被认为是描述物理世界的基本框架，这就是所谓的机械观（mechanical worldview）。在那段时间里，物理学家们曾经试图把物理学的各个分支尽可能地约化为力学。很显然，在那样一个以机械观为主导的时期里，质量既然是力学中的不可约概念，自然也就成为了整个物理学中的不可约概念。不可约概念顾名思义，就是不需要也不能够约化为更基本的概念的，因此有关质量起源的研究在那个时期是基本不存在的①。

但是到了19世纪末的时候，试图把物理学的各个分支约化为力学的努力遭到了很大的挫折。这种挫折首先来自于电磁理论。大家知道，电磁理论预言了电磁波。按照机械观，波的传播必然有相应的介质。但电磁波是在什么介质中传播的呢？却是谁也不知道。尽管如此，物理学家们还是按照机械观的思路假设了这种介质的存在，并称之为“以太”（aether）。但不幸的是，所有试图为以太构筑机械模型的努力全都在实验面前遭遇了滑铁卢。在那段最终催生了狭义相对论的物理学阵痛期里，许多物理学家艰难地试图调和着实验与机械以太模型之间的矛盾。但与那些挽救机械观的努力同时，一种与机械观截然相反的思路也萌发了起来，那便是电磁观（electromagnetic worldview）。电磁观的思路是：物理学上并没有什么先验的理由要求我们用力学的框架来描述自然，机械观的产生只不过是因为力学在很长一个时期里是发展最为成熟的物理学分支而已，现在电磁理论也发展到了不亚于力学的成熟程度，既然无法把电磁理论约化为力学，那何不反过来把力学约化为电磁理论呢？

要想把力学约化为电磁理论，一个很关键的步骤就是把力学中的不可约概念——质量——约化为电磁概念，这是物理学家们研究质量起源的第

① 这里有一个著名的例外是马赫（Ernst Mach，1838—1916年），他对牛顿绝对时空观的批判性思考启示了这样一种观念，那就是一个物体的质量（惯性）起源于宇宙中其他星体的作用。马赫的想法曾对爱因斯坦产生过影响，并且直到现在还有一些物理学家在研究，但它与广义相对论的定量结果及对惯性各向异性的测量结果并不相符。因此我们不把它列为有关质量起源的具体理论。

一种定量尝试。由于当时对物质的微观结构还知之甚少，1897 年由汤姆逊(Joseph John Thomson，1856—1940 年)所发现的电子是当时所知的唯一的基本粒子，因此将质量约化为电磁概念的努力就集中体现在了对电子的研究上，由此产生了物理史上昙花一现的经典电子论(classical electron theory)。

## 四、经典电子论

经典电子论最著名的人物是荷兰物理学家洛伦兹(Hendrik Lorentz，1853—1928 年)，他是一位经典物理学的大师。在相对论诞生之前的那几年里，洛伦兹虽已年届半百，却依然才思敏捷。1904 年，洛伦兹发表了一篇题为《任意亚光速运动系统中的电磁现象》(*Electromagnetic Phenomena in a System Moving with Any Velocity Less than that of Light*)的文章。在这篇文章中他运用自己此前几年在研究运动系统的电磁理论时所提出的包括长度收缩(length contraction)、局域时间(local time)在内的一系列假设，计算了具有均匀面电荷分布的运动电子的电磁动量，由此得到电子的横质量 $m_T$ 与纵质量 $m_L$ 分别为(这里用的是高斯单位制)[①]：

$$m_T = \frac{2}{3}\frac{e^2}{Rc^2}\gamma, \quad m_L = \frac{2}{3}\frac{e^2}{Rc^2}\gamma^3$$

其中 $e$ 为电子的电荷，$R$ 为电子在静止参照系中的半径，$c$ 为光速，$\gamma = (1-v^2/c^2)^{-1/2}$。撇开系数不论，洛伦兹这两个结果所包含的质量与速度的关系与后来的狭义相对论完全相同。

但洛伦兹的文章刚一发表就遭到了经典电子论的另一位主要人物亚伯拉罕(Max Abraham，1875—1922 年)的批评。亚伯拉罕指出，质量除了像洛伦兹

① 洛伦兹所用的质量定义是 $m(\mathrm{d}\boldsymbol{v}/\mathrm{d}t) = \mathrm{d}\boldsymbol{p}/\mathrm{d}t$，“横质量”与“纵质量”分别对应于 $\boldsymbol{v}$ 与 $\mathrm{d}\boldsymbol{v}/\mathrm{d}t$ 垂直及平行这两种特殊情况。

那样通过动量来定义，还应该可以通过能量来定义。比方说纵质量可以定义为 $m_L=(1/v)(dE/dv)$①。但简单的计算表明，用这种方法得到的质量与洛伦兹的结果完全不同。

这说明洛伦兹的电子论是有缺陷的。那么缺陷在哪里呢？亚伯拉罕认为是洛伦兹的计算忽略了为平衡电子内部各电荷元之间的相互排斥所必需的张力。没有那样的张力，洛伦兹的电子会在各电荷元的相互排斥下土崩瓦解②。除亚伯拉罕外，另一位经典物理学大师庞加莱（Henri Poincaré，1854—1912年）也注意到了洛伦兹电子论的这一问题。庞加莱与洛伦兹是爱因斯坦之前在定量结果上最接近狭义相对论的物理学家。不过比较而言，洛伦兹的工作更为直接，为了调和以太理论与实验的矛盾，他提出了许多具体的假设，而庞加莱往往是在从美学与哲学角度审视洛伦兹及其他人的工作时对那些工作进行修饰及完善。这也很符合这两人的特点，洛伦兹是一位第一流的工作型物理学家（working physicist），而庞加莱既是第一流的数学及物理学家，又是第一流的科学哲学家。在 1904 年至 1906 年间，庞加莱亲自对洛伦兹电子论进行了研究，并定量地引进了为维持电荷平衡所需的张力，这种张力因此而被称为庞加莱张力（Poincaré stress）。在庞加莱工作的基础上，1911 年，即在爱因斯坦与闵科夫斯基（Hermann Minkowski，1864—1909 年）建立了狭义相对论的数学框架之后，德国物理学家冯·劳厄（Max von Laue，1879—1960 年）证明了带有庞加莱张力的电子的能量动量具有正确的洛伦兹变换规律。

下面我们用现代语言来简单叙述一下经典电子论有关电子结构的这些主要结果。按照狭义相对论中最常用的约定，我们引进两个惯性参照系：$S$ 与

① 当时还没有爱因斯坦的质能关系式，亚伯拉罕的这一关系式是一个简单的力学关系式，读者不妨自行推导一下。

② 如上所述，亚伯拉罕也是经典电子论的代表人物，有读者可能会问，他自己的电子模型又如何呢？与洛伦兹不同，亚伯拉罕所用的是一个绝对刚性的电子模型，因此在他的模型中不需要引进对能量有贡献的张力。他的模型一度曾被认为比洛伦兹的模型更符合实验，但那实验——即德国物理学家考夫曼（Walter Kaufmann，1871—1947 年）的实验——后来被证实是有缺陷的。

$S'$，$S'$相对于 $S$ 沿 $x$ 轴以速度 $v$ 运动。假定电子在 $S$ 系中静止，则在 $S'$系中电子的动量为

$$p'^{\mu}=\int_{t'=0}T'^{0\mu}(x'^{\xi})\mathrm{d}^3x'=L^0_{\alpha}L^{\mu}_{\beta}\int T^{\alpha\beta}(x^{\xi})\mathrm{d}^3x'$$

其中 $T$ 为电子的总能量动量张量，$L$ 为洛伦兹变换矩阵。由于 $S$ 系中 $T^{\alpha\beta}$ 与 $t$ 无关，考虑到

$$\int T^{\alpha\beta}(x^{\xi})\mathrm{d}^3x'=\int T^{\alpha\beta}(\gamma x',y',z')\mathrm{d}^3x'=\gamma^{-1}\int T^{\alpha\beta}(x^{\xi})\mathrm{d}^3x$$

上式可改写为

$$p'^{\mu}=\gamma^{-1}L^0_{\alpha}L^{\mu}_{\beta}\int T^{\alpha\beta}(x^{\xi})\mathrm{d}^3x$$

由此得到电子的能量与动量分别为（有兴趣的读者可试着自行证明一下）

$$E=p'^0=\gamma m+\gamma^{-1}L^0_iL^0_j\int T^{ij}(x^{\xi})\mathrm{d}^3x$$

$$p=p'^1=\gamma vm+\gamma^{-1}L^0_iL^1_j\int T^{ij}(x^{\xi})\mathrm{d}^3x$$

这里 $i,j$ 的取值范围为空间指标 1，2，3，$m=\int T^{00}(x^{\xi})\mathrm{d}^3x$，为了简化结果，我们取 $c=1$。显然，由这两个式子的第一项所给出的能量动量是狭义相对论所需要的，而洛伦兹电子论的问题就在于当 $T^{\mu\nu}$只包含纯电磁能量动量张量 $T^{\mu\nu}_{\mathrm{EM}}$ 时这两个式子的第二项非零①。

那么庞加莱张力为什么能避免洛伦兹电子论的这一问题呢？关键在于引进庞加莱张力后电子才成为一个满足力密度 $f^{\mu}=\partial_{\nu}T^{\mu\nu}=0$ 的孤立平衡体系。在电子静止系 $S$ 中 $T^{\mu\nu}$不含时间，因此$\partial_jT^{ij}=0$。由此可以得到一个很有用的关系式（请读者自行证明）：$\partial_k(T^{ik}x^j)=T^{ij}$。对这个式子做体积分，注意到左边的积分为零，便可得到

① 有兴趣的读者可以进一步证明这样一些结果：(1)对于球对称均匀面电荷分布，$\int T^{00}_{\mathrm{EM}}(x^{\xi})\mathrm{d}^3x=(1/2)e^2/R$；(2)对于任意球对称电荷分布，$\int T^{ij}_{\mathrm{EM}}(x^{\xi})\mathrm{d}^3x=(1/3)\int T^{00}_{\mathrm{EM}}(x^{\xi})\mathrm{d}^3x$；(3)由 1 和 2 证明洛伦兹有关 $m_{\mathrm{T}}$ 与 $m_{\mathrm{L}}$ 的公式；(4)证实亚伯拉罕对洛伦兹的批评，即用 $m_{\mathrm{L}}=(1/v)(\mathrm{d}E/\mathrm{d}v)$ 定义的质量与洛伦兹的结果不同。

$$\int T^{ij}(x^{\xi})\mathrm{d}^3x = 0$$

这个结果被称为冯·劳厄定理(von Laue's theorem),它表明我们上面给出的电子能量动量表达式中的第二项为零。因此庞加莱张力的引进非常漂亮地保证了电子能量动量的协变性。

至此,经过洛伦兹,庞加莱,冯·劳厄等人的工作,经典电子论似乎达到了一个颇为优美的境界,既维持了电子的稳定性,又满足了能量动量的协变性。但事实上,在这一系列工作完成时经典电子论对电子结构的描述已经处在了一个看似完善,实则没落的境地。这其中的一个原因便是那个"非常漂亮地"保证了电子能量动量协变性的庞加莱张力。这个张力究竟是什么?我们几乎一无所知。更糟糕的是,若真的完全一无所知倒也罢了,我们却偏偏还知道一点,那就是庞加莱张力必须是非电磁起源的(因为它的作用是抗衡电磁相互作用),而这恰恰是对电磁观的一个沉重打击。

就这样,试图把质量约化为纯电磁概念的努力由于必须引进非电磁起源的庞加莱张力而化为了泡影。但这对于很快到来的经典电子论及电磁观的整体没落来说还只是一个很次要的原因。

## 五、量子电动力学

经典电子论的没落是物理学史上最富宿命色彩的事件。这一宿命的由来是因为电子发现得太晚,而量子理论又出现得太早,这就注定了夹在其间,因"电子"而始、逢"量子"而终的经典电子论只能有一个昙花一现的命运①。为它陪葬而终还有建立在经典电磁理论基础上的整个电磁观。

① 当然,这样的说法对历史作了一定的简化。确切地讲,经典电子论的出现实际上略早于电子的发现,而类似于经典电子论的电子结构研究在量子理论之后仍间或地有一些物理学家在做,不过那些研究大都已不能完全归于经典电子论的范畴。另一方面,经典电子论所包含的电子结构以外的东西,比如从物质的微观——但非量子的——电磁结构出发研究宏观电磁及光学性质的方法,直到今天仍可以在一些经典电磁学的教材中找到踪迹。但总体来说,经典电子论随着量子理论的兴盛而没落的大趋势仍是显而易见的。

量子理论对经典物理学的冲击是全方位的，足可写成一部壮丽的史诗。就经典电子论中有关电子结构的部分而言，对这种冲击最简单的启发性描述来自于所谓的不确定原理(uncertainty principle)。如我们在第四节中看到的，经典电子论给出的电子质量——除去一个与电荷分布有关的数量级为 1 的因子——约为 $e^2/Rc^2$。由此可以很容易地估算出 $R \sim 10^{-15}$ 米(感兴趣的读者请自行验证一下)。这被称为电子的经典半径。但是从不确定原理的角度看，对电子的空间定位精度只能达到电子的康普顿波长 $h/mc \sim R/\alpha \sim 10^{-12}$ 米的量级(其中 $\alpha \approx 1/137$ 为精细结构常数)，把电子视为经典电荷分布的做法只有在空间尺度远大于这一量级的情形下才适用。由于电子的经典半径远远小于这一尺度，这表明经典电子论并不适用于描述电子的结构。建立在经典电子论基础上的电子质量计算也因此而失去了理论基础①。

但是经典电子论对电子质量的计算虽然随着量子理论的出现而丧失了理论基础，那种计算所体现的相互作用对电子质量具有贡献的思想却是合理的，并在量子理论中得到了保留。这种贡献被称为电子自能(electron self energy)。在量子理论基础上对电子自能的计算最早是由瑞典物理学家沃勒(Ivar Waller，1898—1991 年)于 1930 年在单电子狄拉克理论的基础上给出的，结果随虚光子动量的平方而发散。1934 年奥地利裔美国物理学家韦斯科夫(Victor Weisskopf，1908—2002 年)计算了狄拉克空穴理论(hole theory)下的电子自能，结果发现其发散速度比沃勒给出的慢得多，只随虚光子动量的对数而发散②。撇开当时那些计算所具有的诸多缺陷不论，韦斯科夫的这一结果在定性上是与现代量子场论一致的。

按照现代量子场论，相互作用对电子自能的贡献可以用对电子传播子产生贡献的单粒子不可约图(one-particle irreducible diagrams)来描述，其中主要部分

---

① 经典电子论对电子的描述不仅与量子力学不符，在电子自旋发现之后，试图在经典电子模型中加入电子自旋的努力与狭义相对论也产生了矛盾，可谓腹背受敌。

② 韦斯科夫的计算包含了一个符号错误，但很快被弗里(Wendell H. Furry，1907—1984 年)和卡尔森(Frank Carlson)所纠正。

来自由量子电动力学(Quantum Electrodynamics, QED)所描述的电磁自能,而电磁自能中最简单的贡献则来自于如图 7 所示的单圈图。幸运的是,由于量子电动力学的耦合常数在所有实验所及的能区都很小,因此这个最简单的单圈图的贡献在整个电子自能中占了主要部分①。

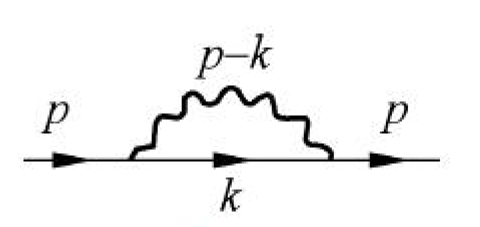


图 7 最简单的电子自能图

对这一单圈图的计算在任何一本量子场论教材中都有详细介绍,其结果为 $\delta m \sim \alpha m \ln(\Lambda/m)$,其中 $m$ 为出现在量子电动力学拉氏量中的电子质量参数,被称为裸质量(bare mass),$\Lambda$ 为虚光子动量的截断(cut-off)能标。如果我们把量子电动力学的适用范围无限外推,允许虚光子具有任意大的动量,则 $\delta m$ 将趋于无穷,这便是自 20 世纪三四十年代起困扰物理学界几十年之久的量子场论发散困难的一个例子。

量子场论中的发散困难,究其根本是由所谓的点粒子模型引起的。这种发散具有相当的普遍性,不单单出现在量子场论中。将经典电子论运用于点电子模型同样会出现发散,这一点从经典电子论的电子质量公式 $m \sim e^2/Rc^2$ 中可以清楚地看到:当电子半径 $R$ 趋于零时质量 $m$ 趋于无穷。经典电子论通过引进电子的有限半径(从而放弃点粒子模型)免除了这一发散,但伴随而来的庞加莱张力、电荷分布等概念却在很大程度上使电子丧失了基本粒子应有的简单性②。这种简单性虽没有先验的理由,但毫无疑问是人们引进基本

---

① 量子场论的微扰展开式有许多微妙的地方。以量子电动力学为例,尽管其耦合常数 $\alpha$ 很小,从而 $n$ 圈图的贡献受到 $\alpha^n$ 的抑制,但另一方面,随着圈数的增加,不等价 $n$ 圈图的数目也在增加,其趋势约为 $n!$(这当然只是非常粗略的说法,圈图的确切数目与相互作用的具体形式有关,且其中还有符号问题,综合的结果非常复杂)。当 $n$ 接近或大于 $1/\alpha$ 时,圈图数量的增加将抵消由弱耦合所带来的减弱因子 $\alpha^n$ 的影响,因此量子电动力学的微扰展开式并不收敛,这一点最早是由英裔美国物理学家戴森(Freeman Dyson)于 1951 年给出的。有鉴于此,所谓单圈图的贡献占了主要部分其实是从渐近级数的意义上说的。

② 顺便提一下,庞加莱张力带来的困难除了我们在第四节中提到的非电磁起源外,还有一个更严重的,那就是由庞加莱张力所维持的电子结构虽然具有静态的平衡,却是不稳定的,在细微的扰动下就会土崩瓦解(类似于爱因斯坦的静态宇宙模型)。这是 1922 年由意大利物理学家费米(Enrico Fermi)所证明的。

粒子这一概念时怀有的一种美学上的期待，正如狄拉克所说："电子太简单，支配其结构的定律根本不应该成为问题。"经典电子论将质量约化为电磁概念的努力即便在其他方面都成功了，其意义也将由于引进电子半径这一额外参数及庞加莱张力、电荷分布等额外假设而大为失色。从这一角度上讲，量子电动力学在概念约化上比经典电子论显得更为彻底，因为在量子电动力学的拉氏量中不含有任何与基本粒子结构有关的几何参数。基本粒子在量子场论中是以点粒子的形式出现的，虽然这并不意味着它们不具有唯象意义上的等效结构，但所有那些结构都是作为理论的结果而不是如经典电子论中那样作为额外假设而出现的，这是除与狭义相对论及量子理论同时兼容，与实验高度相符之外，建立在点粒子模型基础上的量子场论又一个明显优于经典电子论的地方。

至于由此产生的发散困难，在 20 世纪 70 年代之后随着重整化(renomalization)方法的成熟而得到了较为系统的解决。不过尽管人们对重整化方法在数学计算及物理意义的理解上都已相当成熟，发散性的出现在很多物理学家眼里仍基本消除了传统量子场论成为所谓"终极理论"(theory of everything)的可能性，这是后话。

## 六、质量电磁起源的破灭

既然量子电动力学与经典电子论一样具有电子自能，那它能否代替经典电子论实现后者没能实现的把质量完全约化为电磁概念的梦想呢？很可惜，答案是否定的。

这可以从两方面看出来。

首先，从 $\delta m \sim \alpha m \ln(\Lambda/m)$ 中可以看到，由电磁自能产生的质量修正 $\delta m$ 与裸质量 $m$ 的比值为 $\alpha \ln(\Lambda/m)$。由于 $\alpha \approx 1/137$ 是一个比较小的数目，$\ln(\Lambda/m)$ 又是一个增长极其缓慢的函数，因此对于任何普朗克能标以下的截断，$\ln(\Lambda/m)$ 都是一个比较小的数目(特别是，这一数目小于 1)。这意味着由

电磁自能产生的质量修正是比较小的——比裸质量更小①。

另一方面，即便我们一厢情愿地把量子电动力学的适用范围延伸到比普朗克能标还高得多的能区，从而使 $\delta m$ 变得很大，把质量完全约化为电磁概念的梦想依然无法实现。因为电子的电磁自能还有一个很要命的特点，那就是 $\delta m \propto m$。这表明，无论把截断能标取得多大，如果裸质量为零，电子的电磁自能也将为零。因此，为了解释电子质量，裸质量不能为零，而裸质量作为量子电动力学拉氏量中的参数，在量子电动力学的范围之内是无法约化的，从而终结了在量子电动力学中把质量完全约化为电磁概念的梦想。

有的读者可能会问：电磁自能既然是由电磁相互作用引起的，理应只与电荷有关，为什么却会正比于裸质量呢？这其中的奥妙在于对称性。量子电动力学的拉氏量：

$$L = -\frac{1}{4}F^{\mu\nu}F_{\mu\nu} + \bar{\psi}(i\gamma^{\mu}\partial_{\mu} - m)\psi - e\bar{\psi}\gamma^{\mu}A_{\mu}\psi$$

在 $m=0$ 时具有一种额外的对称性，即在 $\psi \to e^{i\alpha\gamma^5}\psi$ 下不变(请有兴趣的读者自行证明)。这种对称性被称为手征对称性(chiral symmetry)，它表明在 $m=0$ 的情形下电子的左右手征态：

$$\psi_L = \frac{1-\gamma^5}{2}\psi, \quad \psi_R = \frac{1+\gamma^5}{2}\psi$$

不会互相耦合。另一方面，(读者可以很容易地证明)电子的质量项

$$m\bar{\psi}\psi = m\,\bar{\psi}_L\psi_R + m\,\bar{\psi}_R\psi_L$$

却是一个电子左右手征态相互耦合，从而破坏手征对称性的项。这样的项在电子的裸质量不存在——从而量子电动力学的拉氏量具有手征对称性——的情况下将被手征对称性所禁止，不可能出现在任何微扰修正中。因此 $\delta m \sim$

① 当我们谈到截断的时候，有一点需要提醒读者注意，那就是对于像电子自能这样对截断能标相对敏感的物理量，只计算截断能标以下的贡献显然是不完整的，那么来自截断能标以上的贡献有多少呢？答案是与适用于截断能标以上的理论的具体形式有关。如果那个理论本身也有截断，我们还必须关心来自那个截断能标以上的贡献。物理学家们的期望是，我们最终将会有一个有限的理论，那时我们就不需要用截断来遮遮掩掩了。

$\alpha m\ln(\Lambda/m)$这一结果的出现是很自然的①。

至此我们看到，试图把质量完全归因于电磁相互作用的想法在量子理论中彻底地破灭了。电磁质量即便在像电子这样质量最小——从某种意义上讲也最为纯粹——的带电粒子的质量中也只占一个不大的比例，在其他粒子——尤其是那些不带电荷的基本粒子——中就更甭提了。

很显然，质量的主要来源必须到别处去寻找。

## 七、对称性自发破缺

质量的电磁起源破灭后，质量起源问题沉寂了很长一段时间。但物理学本身的前进步伐并未因此而停顿。物理学家们手头有大量的观测数据需要分析和解释，同时理论体系本身也有大量的问题亟待解决。对现代物理学的发展来说，这些具体或细节问题是远比解决像质量起源那样的本原问题更重要的动力。另一方面，现代物理学在研究这些具体或细节问题中逐渐积累起来的智慧与洞见，又常常会为更深入地探求本原问题提供新的思路。这是现代物理学的卓越之处，也是它没有像那些只注重于深奥的本原问题，却对细节不屑一顾的其他尝试那样流于肤浅的重要原因。

物理学再次回到质量起源问题是在 20 世纪 60 年代。

在 20 世纪 60 年代初的时候，物理学家们在对基本粒子的研究中已经发现了许多对称性。对称性在物理学中一直有着重要地位，不仅由于其优美的形式与某些物理学家对自然规律的美学追求十分吻合，更重要的是因为它们不仅中看，而且中用，有一种穿透复杂性的力量。即便在对一个物理体系的动力学行为还缺乏透彻理解的情况下，对称性也往往具有令人瞩目的预言能力。这最后

① 这从简单的量纲分析就可以看出：$\delta m$ 的形式为 $mf(\Lambda/m)$，而从费恩曼图所对应的积分的形式可知其相对于 $\Lambda$ 的渐近形式 $f(x)$ 只能是对数或以正负整数为幂次的幂函数，这其中只有 $f(x)=\ln(x)$ 可以使 $\delta m$ 既在 $\Lambda\to\infty$ 时发散，又在 $m\to0$ 时为零。

一点在20世纪五六十年代的粒子物理研究中具有极大的吸引力，因为当时人们对基本粒子相互作用的动力学机制还知之甚少，而且对在很大程度上为研究基本粒子相互作用而发展起来的量子场论产生了很深的怀疑。在这种情况下，许多物理学家对对称性寄予了厚望，希望通过它们来窥视大自然在这一层次上的奥秘。

但不幸的是，当时所发现的许多对称性却被证明只在近似的情况下才成立，比如同位旋对称性。如何理解这种近似的对称性呢？当时有一种猜测，认为近似对称性是（严格）对称性自发破缺的产物。

所谓对称性自发破缺（spontaneous symmetry breaking），指的是这样一种情形：即一个物理体系的拉氏量具有某种对称性，而基态却不具有该对称性。换句话说，体系的基态破缺了运动方程所具有的对称性。这种对称性自发破缺的概念最早是出现在凝聚态物理中的，20世纪60年代被日裔美国物理学家南部阳一郎（Yoichiro Nambu，1921— ）和意大利物理学家约纳-拉西尼奥（Giovanni Jona-Lasinio，1932— ）引进到量子场论中。在量子场论中，体系的基态是真空态，因此对称性自发破缺表现为体系拉氏量所具有的对称性被真空态所破缺。

有的读者可能会问：一个物理体系的真空态是由拉氏量所确定的，为什么会不具有拉氏量所具有的对称性呢？这其中的奥秘在于许多物理体系具有简并的真空态，如果我们把所有这些简并的真空态视为一个集合，它的确与拉氏量具有同样的对称性。但物理体系的实际真空态只是该集合中的一个态，这个态往往不具有整个集合所具有的对称性，这就造成了对称性的破缺——也就是我们所说的对称性自发破缺①。

但是把近似对称性归因于对称性自发破缺的想法在1961年遭到了致命

① 学过量子力学的读者可能会进一步问：如果一个量子体系的基态是简并的，那么体系的物理基态难道不应该是这些简并态的某种量子叠加吗？这种量子叠加——如我们在量子力学中所见到的——往往不仅会破除原有的基态简并性，并且使真正的基态具有与原先简并基态的**集合**相同的对称性。在这种情况下，对称性自发破缺岂不是不存在了？这是一个非常好的问题，答案是：对于**有限体系**来说情况确实会如此（除非有什么原因——比如对称性——禁止简并基态间的相互耦合）。但在量子场论中通常假定体系的空间体积趋于无穷，这时不同真空态之间的相互耦合趋于零，严格的对称性自发破缺只发生在这种情形下。

的打击。那一年由英国物理学家戈德斯通（Jeffrey Goldstone，1933— ）提出并在稍后与巴基斯坦物理学家萨拉姆（Abdus Salam，1926—1996年）及美国物理学家温伯格（Steven Weinberg，1933— ）一起证明了这样一个命题——被称为戈德斯通定理（Goldstone theorem）：每一个自发破缺的**整体连续对称性**都必然伴随一个无质量标量粒子。这个无质量标量粒子被称为戈德斯通粒子（Goldstone particle）或南部-戈德斯通粒子（Nambu-Goldstone particle）。

为什么会有这样的结果呢？我们来简单地证明一下：

假定一个物理体系的拉氏量中的势函数为 $V(\varphi_a)$ $(a=1,\cdots,N)$，其中 $\varphi_a$ 为标量场（可以是基本的也可以是复合的）。显然，该体系的真空态满足 $\partial V/\partial\varphi_a=0$（为避免符号繁复，我们略去了对真空的标记），而标量粒子的质量（平方）由 $\partial^2 V/\partial\varphi_a\partial\varphi_b$ 在真空态上的本征值给出。现在考虑对真空态 $\varphi_a$ 作一个无穷小连续对称变换 $\varphi_a\to\varphi_a+\varepsilon\Delta_a(\varphi)$（其中 $\varepsilon$ 为无穷小参数）。由于 $V(\varphi_a)$ 在这一变换下不变（请读者想一想这是为什么），因此有 $\Delta_a(\varphi)(\partial V/\partial\varphi_a)=0$（对相同指标求和，下同）。将这一表达式对 $\varphi_b$ 作一次导数，并注意到真空所满足的条件，可得（请读者自行证明）：

$$\Delta_a(\varphi)\,\frac{\partial^2 V}{\partial\varphi_a\partial\varphi_b}=0$$

由上式可以看到，每一个 $\Delta_a(\varphi)\neq 0$ 的连续对称变换都对应于 $\partial^2 V/\partial\varphi_a\partial\varphi_b$ 的一个本征值为零的本征态，从而也就对应于一个无质量标量粒子。而另一方面，$\Delta_a(\varphi)\neq 0$ 的连续对称变换所对应的正是那些不能使真空态不变——从而被真空态所破缺（即自发破缺）——的连续对称性。这就证明了每一个自发破缺的整体连续对称性都必然伴随一个无质量标量粒子，即戈德斯通粒子。这正是戈德斯通定理。① （请读者思考一下，戈德斯通定理中的“整体”二字体现

① 戈德斯通定理也可以从几何上来理解。$V=V(\varphi_a)$ $(a=1,\cdots,N)$ 可以看成是一个 $N$ 维曲面，真空态对应于该曲面的一个极小值点，而该点处每一个独立的平坦方向（即二阶导数为零的方向）对应于一个无质量标量粒子。另一方面，每一个这种独立的平坦方向对应于一个可以使真空态移到邻近点的连续对称变换。这种连续对称变换所表示的正是被真空态所破缺的对称性。这就表明无质量标量粒子与这种自发破缺的对称性一一对应。另外再补充一点：南部阳一郎曾在1960年提出过类似于戈德斯通定理的想法，但未引起足够重视。

在证明的什么地方?)由于自发破缺的整体连续对称性的数目等于这些对称性的生成元的数目,因此戈德斯通定理也表明了戈德斯通粒子的数目等于自发破缺的整体连续对称性生成元的数目。举个例子来说,SU(2)对称性具有三个生成元,若完全破缺,就会产生三个戈德斯通粒子;若破缺为U(1),则只产生两个戈德斯通粒子(因为有一个生成元未破缺)。进一步的分析还表明,戈德斯通粒子与那些自发破缺的整体连续对称性所对应的荷——关于荷,请读者回忆一下诺特定理(Noether theorem)——具有相同的宇称及内禀量子数。

当然,严格讲,上面的证明只是在所谓经典层次上的证明,而没有考虑量子修正。那么考虑了量子修正后,戈德斯通定理是否仍成立呢?答案是肯定的,而且证明也基本一样,只需用包含量子修正的所谓量子有效势 $V_{\text{eff}}$ 取代经典拉氏量中的势函数 $V$ 即可①。

由戈德斯通等人证明的这一结果为什么会对把近似对称性归因于对称性自发破缺的想法造成致命打击呢?原因很简单,那就是近似对称性中的某一些——比如同位旋对称性——正是整体连续对称性,如果它们的近似性果真源自对称性自发破缺,那就应该存在相应的无质量标量粒子。但我们从未在实验上观测到任何这样的粒子。因此对称性自发破缺的想法在粒子物理学中由于牵涉到无质量粒子而陷入了困境。

---

① 这里有一个很有意思的问题,那就是既然真正的对称性自发破缺是由量子有效势 $V_{\text{eff}}$ 而非经典势函数 $V$ 所决定的,那么在经典势函数 $V$ 不具有简并真空态(从而不会产生对称性自发破缺)的情况下,是否有可能通过体现在有效势 $V_{\text{eff}}$ 中的纯量子效应产生对称性自发破缺呢?答案是肯定的。如果哪位读者**独立地**想到了这个问题,那么祝贺你了,这说明你有非常敏锐的物理思维能力。如果你同时还具有第一流的理论基础,并且早生几十年的话,就有可能作出一个非常重大的理论发现,那便是1973年由美国物理学家科尔曼(Sidney Coleman,1937—2007年)与温伯格(Erick Weinberg,1947— )所发现的如今被称为科尔曼-温伯格机制(Coleman-Weinberg mechanism)的对称性破缺机制。

## 八、从希格斯机制到电弱统一理论

无独有偶，粒子物理学中产生于20世纪五六十年代的另一个很高明的想法也受到了无质量粒子的困扰，那个想法是1954年由杨振宁（1922— ）和米尔斯（Robert Mills，1927—1999年）提出的，现在被称为杨-米尔斯理论（Yang-Mills theory）。这是一种所谓的定域"非阿贝尔规范理论"（non-Abelian gauge theory），是对像量子电动力学那样的定域"阿贝尔规范理论"（Abelian gauge theory）的推广①，具体的区别是以非阿贝尔规范对称性取代了量子电动力学所具有的阿贝尔规范对称性——即U(1)规范对称性。提出这种理论最初的动机是企图用它来描述同位旋对称性。但这一企图立刻就遇到了一个很大的困难，那便是这种理论所具有的定域规范对称性会无可避免地导致无质量的矢量粒子（被称为规范粒子，类似于量子电动力学中的光子），而在现实中，除光子外我们从未在实验上观测到任何这样的粒子。

就这样，杨-米尔斯理论与对称性自发破缺这两个出色的想法先后搁浅了，推根溯源，都是无质量粒子惹的祸。但如果我们仔细研究一下这对"难兄难弟"的病根，就会发现两者竟然像是互为解药！对称性自发破缺的问题出在哪里呢？出在**整体**连续对称性上；而杨-米尔斯理论的问题又出在哪里呢？出在**定域**规范对称性（那是一种特殊的定域连续对称性）上。如果我们把这两者放在一起，让对称性自发破缺干掉那些产生无质量矢量粒子的定域规范对称性，杨-米尔斯理论不就可以摆脱困境了吗？更妙的是，由于杨-米尔斯理论中的对称性不是整体而是定域的，戈德斯通定理将不适用于这种对称性的自发破缺，这样一来说不定那些可恶的戈德斯通粒子也会消失，那岂不是两全其美？世界上会有这么好的事吗？还真的有。

---

① 一般来说，粒子物理学中的规范对称性指的就是"定域"规范对称性。不过在本节中，为突出"定域"所起的作用，我们有时会特意注明。

最早明确指出这一点的是美国凝聚态物理学家安德森(Philip Warren Anderson, 1923— )。对于安德森来说,戈德斯通定理显然不可能是普遍成立的,因为当时的凝聚态物理学家们已经知道,超导体就是一个连续对称性——U(1)对称性——自发破缺的体系,但在这一破缺的过程中并没有产生无质量的戈德斯通粒子。安德森并且很正确地意识到了U(1)对称性的**定域**特点是使戈德斯通定理失效的关键。由于并非只有定域U(1)对称性具有定域特点,事实上所有杨-米尔斯理论也都具有这一特点。因此安德森在1963年猜测道:"戈德斯通的零质量困难并不是一个严重的困难,因为我们很可能可以用一个相应的杨-米尔斯零质量问题来消去它。"安德森的想法得到了一些物理学家的认同,但也有人认为这种凝聚态物理的类比不能应用到相对论量子场论中。

这种怀疑很快就被推翻了。1964年,英国物理学家希格斯(Peter Higgs, 1929— )、比利时物理学家英格勒特(François Englert, 1932— )与布罗特(Robert Brout, 1928— )等几乎同时证实了安德森的想法。这便是描述规范对称性自发破缺的著名的希格斯机制(Higgs mechanism),它一方面消除了无质量的戈德斯通粒子,另一方面则使规范粒子获得了质量①。

不过希格斯等人的漂亮工作并没有引起即刻的轰动。希格斯就这一工作所写的两篇短文中的第二篇甚至一度遭到了退稿,理由是"与物理世界没有明显关系"。这一退稿理由使希格斯深感不快,但也促使他更深入地考虑了理论可能引致的实验结果,并对论文进行了补充。希格斯后来认为,他因遭到退稿

① 用技术性的语言来说,在希格斯机制中对应于戈德斯通粒子的那些自由度可以被定域规范变换所消去(必须注意的是:"定域"二字在这里至关重要,整体的连续变换是不具有这种能力的)。从规范理论的角度讲,这相当于选取了一种被称为幺正规范(unitary gauge)的特殊规范。这种特殊规范的选取造成定域规范对称性的破缺,从而使原本受定域规范对称性所限必须无质量的规范粒子可以获得质量。人们有时把这种机制形象地描述为:规范粒子通过"吃掉"戈德斯通粒子而获得质量。另外要说明的是,这里所介绍的由希格斯等人提出的,被粒子物理标准模型所吸收的其实只是希格斯机制的一种最简单的实现形式——但似乎恰好就是自然界所采用的形式。

而补充的那些内容是人们将希格斯粒子及希格斯机制与他的名字联系在一起的主要原因。

做了这么多背景介绍，现在让我们回到主题——质量的起源——上来。希格斯机制不仅一举“救活”了粒子物理学中对称性自发破缺与杨-米尔斯理论这两个极为出色的想法，而且在救助过程中为我们提供了一种产生质量的新方法，即通过规范对称性的自发破缺，从不带质量项的拉氏量中产生出质量来。不过，由此而获得质量的——如上文及注释所述——只是规范粒子，而规范粒子的质量在宇宙可见物质的质量中只占了微不足道的比例，我们更关心的是在可见物质质量中占主要比例的那些粒子——费米子。

那么，费米子的情况如何呢？1967 年，温伯格和萨拉姆将希格斯机制应用到美国物理学家格拉肖（Sheldon Lee Glashow，1932— ）等人几年前所提出的旨在描述电磁和弱相互作用的 SU(2)×U(1)规范理论中，建立起了所谓的电弱统一理论（electroweak theory）[①]。这一理论与描述强相互作用的量子色动力学（quantum chromodynamics）一起组成了粒子物理的标准模型。在标准模型中，费米子也是通过规范对称性的自发破缺——或者更确切地说，通过电弱统一理论中的规范对称性自发破缺——获得质量的。具体地讲，在标准模型中，费米场 $\psi$ 与希格斯机制中的标量场（也称为希格斯场）$\varphi$ 之间存在所谓的汤川耦合（Yukawa coupling）：$-\lambda\bar{\psi}\psi\varphi$（其中 $\lambda$ 为耦合常数）[②]。由于希格斯场 $\varphi$ 具有非零的真空期待值，因此将这一耦合项相对于真空展开后就会出现形如 $-m\bar{\psi}\psi$ 的费米子质量项。

因此，我们可以说，标准模型中所有基本粒子的质量都来源于电弱统一理

① 电弱统一理论中的规范对称性破缺方式是 SU(2)×U(1)破缺为 U(1)，由此产生的三个戈德斯通粒子通过希格斯机制使四个规范粒子中的三个（即 $W^{\pm}$ 和 Z）获得质量，剩下的一个（即光子）则维持了无质量。

② 更确切地讲，标准模型中的汤川耦合是形如 $-\lambda\bar{\psi}_L\psi_R\varphi$ − h. c. 的项，其中 $\psi$ 为质量本征态（不同于弱本征态），L 与 R 分别代表左右手征部分，h. c. 代表厄密共轭。汤川耦合是费米子场与标量场之间唯一的可重整耦合。

论中的规范对称性自发破缺。这是标准模型对质量起源问题的直接回答。

不过遗憾的是,这一回答却是一个不尽人意的回答。为什么这么说呢?因为这一回答从某种意义上讲与其说是回答了问题,不如说是在转嫁问题——把我们想要理解的基本粒子的质量转嫁给了希格斯场的真空期待值、规范耦合常数以及汤川耦合常数。这其中希格斯场的真空期待值及规范耦合常数与基本粒子——主要是费米子——的种类无关,可以算是具有普适性的,因此将质量向这些参数约化不失为是一种有效的概念约化。但汤川耦合常数则不然,它对于每一种费米子都有一个独立的数值。由于这些参数的存在,标准模型的拉氏量虽然不显含质量参数,但它所包含的与质量直接有关的自由参数的数目却一点也不比原先需要解释的质量参数的数目来得少(事实上还略多一点)。从某种意义上讲,用这种方式来解释质量的起源,就像英国物理学家霍金(Stephen Hawking,1942— )在《时间简史》(*A Brief History of Time*)一书中引述的一位老妇人的"理论"。那位老妇人宣称世界是平面的,由一只大乌龟托着。当被问到那只大乌龟本身站在哪里时,老妇人冷静地回答说:"站在另一只大乌龟的背上。"

因此,希格斯机制及包含希格斯机制的电弱统一理论虽然从许多唯象的方面来衡量是非常成功的,其所体现的把质量与真空的对称性破缺性质联系在一起的思路也极为深刻。但它们作为与对称性破缺有关的特殊机制或模型,本身却没能实现对质量概念的真正约化,从而不能被认为是对质量起源问题令人满意的回答。

## 九、量子色动力学

与戈德斯通、希格斯等人在对称性自发破缺方面的研究几乎同时,物理学家们在研究强相互作用上也取得了重大进展。1961 年,美国物理学家盖尔曼(Murray Gell-Mann,1929— )与以色列物理学家内曼(Yuval Ne'eman,

1925—2006 年)彼此独立地提出了强子分类的 SU(3)模型①。这一模型不仅对当时已知的强子给出了很好的分类,而且还预言了当时尚未发现的粒子,比如 $\Omega^-$ 粒子②。但这一模型有一个显著的缺陷,那就是 SU(3)的基础表示(foundamental representation)似乎不对应于任何已知的粒子。1964 年,盖尔曼与美国物理学家茨威格(George Zweig, 1937— )提出了夸克(quark)模型,将夸克作为 SU(3)基础表示所对应的粒子,强子则被视为是由夸克组成的复合粒子③。

在夸克模型中,为了给出正确的强子性质,夸克必须具有实验上从未发现过的量子数,比如分数电荷,这在当时是令人不安的。对此,盖尔曼也深感困惑,只能用“夸克存在但不是真实的”(they exist but are not real)这样诡异的语言来搪塞。夸克模型的另一个麻烦是,夸克是费米子,而某些强子却似乎包含三个处于同一量子态的夸克,从而违反了泡利不相容原理(Pauli exclusion principle)。关于这一点,1965 年美国物理学家格林伯格(Oscar W. Greenberg, 1932— )、韩国物理学家韩武永(Moo-Young Han, 1934— )和南部阳一郎先后提出了一个解决方案,那就是引进一个新的三值量子数以保证那些夸克具有不同的量子态。南部阳一郎甚至粗略地设想了以这一量子数为基础构造杨-米尔斯理论,但这些工作并未引起重视。1972 年,盖尔曼等人在实验的引导下重新考虑了这一被盖尔曼称之为色荷(color)的新量子数,以及

① 盖尔曼将这一模型称为八正道(eightfold way),这一名称取意于佛教术语,所代表的是 SU(3)分类模型中的八维表象。

② $\Omega^-$ 粒子于 1964 年被发现,它不仅量子数与理论预言完全一致,质量也非常接近理论的预期。

③ 当时盖尔曼是加州理工大学(California Institute of Technology)的教授,茨威格则是该校的研究生,他们虽在同一学校,但提出夸克模型是彼此独立的。夸克这一名称是盖尔曼所取,来自于爱尔兰作家乔伊斯(James Joyce, 1882—1941 年)的小说《芬尼根的守灵夜》(*Finnegans Wake*);茨威格提议的名字也很幽默,是“Aces”——即扑克牌中的“爱斯”。对茨威格来说,十分苦涩的经历是:同样标新立异的理论,盖尔曼的文章应杂志编辑的亲自邀请发表在了欧洲核子中心(CERN)的新杂志《物理快报》(*Physics Letters*)上,而人微言轻的茨威格的文章却遭到拒稿而未能及时发表。茨威格后来转行离开了物理。

以之为基础的杨-米尔斯理论。这一理论被称为了量子色动力学。由于色荷是一个三值量子数，因此量子色动力学的规范群被选为了 SU(3)。

在量子色动力学的发展过程中，20 世纪 60 年代末的一系列所谓“电子-核子深度非弹性散射”(deep-inelastic electron-nucleon scattering)实验起了很大的作用。这些实验不仅证实了核子内部存在着点状结构，而且还显示出这些点状结构之间的相互作用在高能——即近距离——下会变弱。这些点状结构被美国物理学家费恩曼(Richard Feynman，1918—1988 年)称为“部分子”(parton)，它们中的一部分后来被证实就是夸克(另一部分是后面会提到的胶子)，而部分子之间的相互作用在高能——即近距离——下变弱的行为则被称为渐近自由(asymptotic freedom)。渐近自由为实验上从未观测到孤立夸克这一事实提供了一种很好的说明：那就是当夸克彼此远离时，它们之间的相互作用会越来越强，最终从真空中产生出足以中和它们所带色荷的粒子。我们在实验上能够分离出的任何粒子——比如强子——都只能是这种色荷中和之后的产物，而不可能是孤立的夸克①。由于这一原因，渐近自由很快被视为描述夸克相互作用的理论所必须具备的性质。

1973 年，美国物理学家波利策(Hugh David Politzer，1949— )、韦尔切克(Frank Wilczek，1951— )和格娄斯(David Gross，1941— )等人发现杨-米尔斯理论具有渐近自由性质②。在当时已知的所有四维可重整场论中，杨-米尔斯理论是唯一具备这一性质的理论，这对盖尔曼等人提出的量子色动力学是一个很强的支持。那时候，人们对杨-米尔斯理论本身的研究也已取得了系

① 这一点也适用于胶子或任何不处于色单态的粒子组合。不过要注意的是，它的严格数学证明是极其困难的。事实上，它是美国克莱数学研究所(Clay Mathematics Institute)悬赏百万美元征解的七大数学难题之一的“杨-米尔斯与质量隙”(Yang-Mills and Mass Gap)问题的一部分。不过许多物理学家对从数学上严格证明这一点并无太大兴趣，温伯格就曾经表示：“这一点肯定是正确的，因此我和其他一些人一样很乐意把证明留给数学家去做。”

② 波利策等人因此而获得了 2004 年的诺贝尔物理学奖。比他们稍早，荷兰物理学家特·胡夫特也有过同样的发现，可惜没有发表。

统性的进展：1967 年，苏联物理学家法捷耶夫(Ludvig Faddeev，1934— )和波波夫(Victor Popov，1937—1994 年)完成了杨-米尔斯理论的量子化；1971 年，荷兰物理学家特・胡夫特(Gerard't Hooft，1946— )证明了杨-米尔斯理论的可重整性。在这一系列工作的基础上，量子色动力学顺理成章地成为了标准模型中描述强相互作用的基本理论。这一理论中对应于 SU(3)生成元的八个载力子被称为胶子(gluon)，它们都是无质量的。

看到这里，有些读者可能会问：我们是不是离题了？量子色动力学中总共只有两类粒子：胶子与夸克。其中胶子是无质量的，而夸克虽然有质量，但其质量——与标准模型中其他费米子的质量一样——却是由电弱统一理论中的规范对称性自发破缺产生的，与量子色动力学无关。既然如此，量子色动力学与质量起源这一主题又能有什么关系呢？应该说，这是一个很合理的疑问。但量子色动力学的奇妙之处就在于，它形式上异常简洁——一个简简单单的规范群，一个平平常常的耦合常数，差不多就是全部的家当——但内涵却惊人地丰富。它宛如一坛绝世的佳酿，越品就越是回味无穷。在谈论质量起源问题的时候，人们往往把注意力放在希格斯机制及包含希格斯机制的电弱统一理论上——因为希格斯机制在登场伊始就打出了质量产生机制的响亮广告。但事实上我们将会看到，**看似与质量起源问题无关的量子色动力学对这一问题有着非常独特而精彩的回答**，而且从某种意义上讲，这一回答才是标准模型范围内的最佳回答。

我们先来看看量子色动力学的拉氏量：

$$L = -\frac{1}{2}\mathrm{Tr}(G^{\mu\nu}G_{\mu\nu}) + \sum_q \bar{q}(\mathrm{i}\gamma^\mu D_\mu - m_q)q$$

其中 $q$ 为夸克场；$G_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu - \mathrm{i}g[A_\mu, A_\nu]$ 为规范场强；$D_\mu = \partial_\mu - \mathrm{i}gA_\mu$ 为协变导数；$A_\mu$ 为规范势；$m_q$ 为夸克 $q$ 的质量；$g$ 为耦合常数；式中的求和遍及所有的夸克种类。自然界已知的夸克种类——也称为“味”(flavor)——共有六种。其中 u(上夸克)、d(下夸克)、s(奇夸克)被称为轻夸克，质量分别约为 2.3MeV、4.8MeV 和 95MeV；c(粲夸克)、b(底夸克)、t(顶夸克)被称为重夸

克，质量分别约为 1.3GeV、4.2GeV 和 173GeV。这其中轻夸克的质量是在约 2GeV 的能标上定义的，重夸克的质量则是在其自身质量标度上定义的①。这些质量参数本身在标准模型范围内是不能约化的，但由这些夸克所组成的强子的性质，在很大程度上可以由量子色动力学来描述，这其中就包括强子的质量。

在接下来的几节中，我们就来看一下量子色动力学对强子质量的描述，以及这种描述在何种意义上可以被视为是对质量起源问题的回答。

## 十、同位旋与手征对称性

我们知道，可见物质的质量主要来自于质子和中子，其中质子由两个 u 夸克及一个 d 夸克组成，而中子由一个 u 夸克及两个 d 夸克组成。在下面的叙述中，我们将只考虑这两种夸克。由于这两种夸克的质量远小于包括质子和中子在内的任何强子的质量，作为近似，我们先忽略它们的质量。这时量子色动力学的拉氏量为

$$L = -\frac{1}{2}\mathrm{Tr}(G^{\mu\nu}G_{\mu\nu}) + \mathrm{i}\,\bar{u}\gamma^{\mu}D_{\mu}u + \mathrm{i}\,\bar{d}\gamma^{\mu}D_{\mu}d$$

显然（请读者自行验证），这一拉氏量在以下两个整体 SU(2)变换：

$$\psi \to \exp(-\mathrm{i}t^a\theta^a)\psi, \quad \psi \to \exp(-\mathrm{i}\gamma^5 t^a\theta^a)\psi$$

下是不变的。这其中 $\psi=(u,d)^{\mathrm{T}}$，$t^a$ 是 SU(2)的生成元（即泡利矩阵的 1/2）。这两个存在于 u 夸克和 d 夸克之间的对称性分别被称为同位旋对称性与手征对称性（chiral symmetry），记为 $SU(2)_V$ 与 $SU(2)_A$。这其中同位旋对称性 $SU(2)_V$ 只要夸克质量彼此相等（不一定要为零）就存在，而手征对称性 $SU(2)_A$ 只有在夸克质量全都为零时才具有（这一情形因此而被称为手征极限）。这一点与我们在第六节中提到的无质量量子电动力学的手征对称性类

① 补充说明两点：(1)定义夸克质量所用的重整化方案（renormalization scheme）是 $\overline{\mathrm{MS}}$。(2)夸克的“轻”和“重”是相对于量子色动力学中的特征能标 $\Lambda_{\mathrm{QCD}}$（约为 200～300MeV）来区分的。

似。除此之外，这一拉氏量还存在一个显而易见的整体 $U(1)_V$ 对称性，它对应于重子数守恒，与夸克是否有质量，以及质量是否彼此相等都无关。

综合起来，忽略夸克质量的上述拉氏量具有整体 $SU(2)_V \times SU(2)_A \times U(1)_V$ 对称性①。在这些对称性中，同位旋对称性 $SU(2)_V$ 与手征对称性 $SU(2)_A$ 所对应的守恒流分别为

$$V^{\mu a} = \bar{\psi}\gamma^\mu t^a \psi, \quad A^{\mu a} = \bar{\psi}\gamma^\mu \gamma^5 t^a \psi$$

显然，在宇称变换下，$V^{\mu a}$ 是矢量（vector），$A^{\mu a}$ 则是轴矢量（axial vector）。它们对应的荷 $(Q_V)^a = \int V^{0a} \mathrm{d}^3 x$ 与 $(Q_A)^a = \int A^{0a} \mathrm{d}^3 x$ 分别为标量（scalar）及赝标量（pseudoscalar）②。

如果同位旋与手征对称性都是严格的对称性，那么 $(Q_V)^a$ 将生成强子谱中自 20 世纪 60 年代起逐步引导人们发现量子色动力学的同位旋对称性；而 $(Q_A)^a$ 则将生成所谓的手征对称性，它要求每一个强子都伴随有自旋、重子数及质量与之相同，而宇称却相反的粒子——**那样的对称性在强子谱中并未被发现过**。

对此，最容易想到的解释是：由于 u 夸克和 d 夸克实际上并不是无质量的，因此手征对称性本就不可能严格成立。事实上，不仅手征对称性不可能严格成立，由于 u 夸克和 d 夸克的质量彼此不同，连同位旋对称性也不可能严格成立。但是，考虑到 u 夸克和 d 夸克的质量相对于强子质量是如此之小，相应的对称性在强子谱中似乎起码应该近似地存在。对于同位旋对称性来说，情况的确如此（否则就不会有早年那些强子分类模型了）③。但手征对称性却哪

---

① 有读者可能会问：既然有 $U(V)_V$，是不是也有 $U(1)_A$？在经典层次上答案是肯定的，但是在量子世界里，$U(1)_A$ 会被反常（anomaly）所破坏。

② 感兴趣的读者请利用场量的宇称变换性质 $\psi(t,x) \to \gamma^0 \psi(t,-x)$ 自行证明 $V^{\mu a}$ 与 $A^{\mu a}$ 的变换性质 $V^{\mu a}(t,x) \to V_\mu^a(t,-x)$ 与 $A^{\mu a}(t,x) \to -A_\mu^a(t,-x)$。另外要注意的是，这里所说的矢量、轴矢量、标量、赝标量都是依据时空变换性质区分的，与那些量在 SU(2) 内禀空间内的变换性质无关。

③ 由于 s 夸克也是轻夸克，因此我们的讨论可以扩展至包括 s 夸克，这是强子分类中存在 SU(3) 近似对称性的原因——请注意这个 SU(3) 是“味”对称性而不是“色”对称性。不过由于 s 夸克的质量较大，SU(3) 对称性的近似程度远不如 SU(2) 对称性来得高。

怕在近似意义上也根本不存在。举个例子来说，手征对称性要求介子三重态 $\rho(770)$与 $a_1(1260)$互为对称伙伴（请读者自行查验这两组介子的量子数），但实际上这两者的质量分别约为 775MeV 和 1230MeV①，相差悬殊（作为对比，同位旋伙伴的质量差通常都在几个 MeV 以下），连近似的对称性也不存在。

初看起来，事情似乎出了麻烦，但物理学家们却从这一麻烦中找到了一条探究低能量子色动力学的捷径。正所谓“山重水复疑无路，柳暗花明又一村”。

## 十一、手征对称性自发破缺

手征对称性 $SU(2)_A$ 是量子色动力学拉氏量中的（近似）对称性，却在现实世界中完全找不到对应，这究竟是什么原因呢？应该说，要猜测一下是不困难的，因为当时物理学家们已经知道对称性可以自发破缺。如果量子色动力学中的手征对称性是自发破缺的，显然就会出现这种拉氏量具有（近似）手征对称性，现实世界却不并不买账的现象。但是，猜测归猜测，要想在理论上严格证明这一点——哪怕只是在物理学而不是数学的标准下严格证明——却是极其困难的。

有读者可能会问：对称性自发破缺在电弱统一理论中用得好好的，为什么在量子色动力学中却变得“极其困难”了呢？这是因为在电弱统一理论中对称性自发破缺是由人为引进的希格斯场产生的，我们有一定的自由度来选择对称性破缺的方式。但量子色动力学并不包含这种人为引进的希格斯场，因此，在量子色动力学中，整体 $SU(2)_V\times SU(2)_A\times U(1)_V$ 对称性是否自发破缺？如果破缺，是否恰好是手征部分 $SU(2)_A$ 破缺，即破缺到 $SU(2)_V\times U(1)_V$？都只能由理论本身来决定，而不是我们可以擅自假设的，正是这一特

---

① 在强子的命名中，有些带有质量参数，$\rho(770)$与 $a_1(1260)$就是两个例子。细心的读者可能要问：既然如此，这两个介子的质量怎么会是 775MeV 和 1230MeV，而非 770MeV 和 1260MeV 呢？我把这个问题留给读者自己去思考。

点使问题变得“极其困难”①。更麻烦的是，手征对称性的破缺——如果出现的话——乃是一种出现在量子色动力学的强相互作用区域——即低能区域——的现象。对于理论研究来说，这无疑是雪上加霜。

另一方面，**对称性自发破缺的存在与否及具体方式由理论本身所决定，虽然为量子色动力学带来了一个“极其困难”的理论问题，同时却也是它的一个极大的理论优势**。因为电弱统一理论之所以只是对质量起源问题的一个不尽人意的回答，一个很重要的原因就是希格斯场以及它与费米场之间的相互作用——汤川耦合——都是人为引进的，从而都是所谓的自由参数（free parameter）。而量子色动力学没有那种类型的自由参数，因此它与观测之间的对比更为严酷：如果成功，将是极具预言能力的成功，因为自由参数越少，预言能力就越强；但如果失败，也将是无力回天的失败，因为自由参数越少，回旋余地也就越小。

那么量子色动力学究竟能不能实现从 $SU(2)_V \times SU(2)_A \times U(1)_V$ 到 $SU(2)_V \times U(1)_V$ 的对称性自发破缺呢？目前在理论上还是一个待解之谜。1979 年，特·胡夫特通过对规范理论中的反常（anomaly）进行分析，得到了一个结果：即如果所考虑的整体对称性是 $SU(3)_V \times SU(3)_A \times U(1)_V$，那它就必须自发破缺。可惜的是，一来量子色动力学中的 SU(3) 对称性远比 SU(2) 对称性粗糙，二来这一结果并未告诉我们具体哪一部分对称性会自发破缺。1980 年，美国物理学家科尔曼（Sidney Coleman，1937—2007 年）与威顿（Edward Witten，1951—  ）提出了在某些合理的物理条件下，当色的数目 $N_c$ 趋于无穷时，手征对称性必须自发破缺。这一结果虽然抓准了手征对称性，但可惜量子色动力学中色的数目 $N_c$ 不仅不是无穷，而且还很小（$N_c=3$）。1984 年，伊朗裔美国物理学家瓦法（Cumrun Vafa，1960— ）与威顿证明了

① 虽然从实验上观测到的强子谱来看，量子色动力学中的 $SU(2)_V \times SU(2)_A \times U(1)_V$ 对称性几乎肯定是破缺成了 $SU(2)_V \times U(1)_V$（即手征对称性被破缺了），但这并不意味着量子色动力学的真空一定能够实现这一破缺方式。相反，能否实现这一破缺方式在很大程度上可以视为是对量子色动力学的检验。

未被非零夸克质量项所破缺的同位旋对称性(请读者想一想,在现实世界里这一对称性由什么群来表示?)不会自发破缺。可惜这一证明虽然表明特定的同位旋对称性不会自发破缺,却未能对手征对称性是否一定会自发破缺提供说明。

虽然上述理论研究没有一个能够证明量子色动力学中的 $SU(2)_V \times SU(2)_A \times U(1)_V$ 整体对称性必定会自发破缺到 $SU(2)_V \times U(1)_V$,但它们都与这一破缺方式相容这一事实,无疑还是大大增强了人们的信心。在物理学上,严格证明是一种美妙的东西,但有时却可望不可及,物理学家们的工作往往并不总是依赖于它。迄今为止,虽然尚未有人能够给出量子色动力学中手征对称性自发破缺的严格证明,但从这一破缺方式已经得到的大量间接证据来看,它的证明应该只是时间问题。物理学家们更感兴趣的是:如果手征对称性自发破缺,我们可以从中得到什么推论?有关这一点,人们做过不少细致研究。那些研究获得了极大的成功,不仅给出了被称为“手征微扰理论”(chiral perturbation theory)的描述低能量子色动力学的所谓“有效场论”(effective field theory),而且得到了一系列与实验相吻合的漂亮结果。这一切也反过来为手征对称性的自发破缺提供了进一步的间接证据。

下面我们就来看看由手征对称性自发破缺导致的推论之中与质量起源问题有密切关系的部分。

## 十二、赝戈德斯通粒子的质量

我们在第七节中介绍过,对称性自发破缺的最重要的推论之一,是存在无质量的标量粒子,即戈德斯通粒子,它们与自发破缺的对称性所对应的荷具有相同的宇称及内禀量子数。对于手征对称性来说,荷是$(Q_A)^a$,它在时空中是一组赝标量,在内禀空间中则是一个矢量,因此相应的戈德斯通粒子的宇称为负,同位旋则为 1。自然界里满足这些特征的强子中质量最轻的是 π 介子($\pi^-$、$\pi^0$ 和 $\pi^+$)。如果手征对称性是自发破缺的,π 介子就应该是这一破缺所

对应的戈德斯通粒子①。但是，戈德斯通粒子是无质量的，$\pi$ 介子却是有质量的，这一矛盾该如何解决呢？

我们知道，在理想的对称性自发破缺情形下，体系的实际真空态可以是一系列简并真空态中的任何一个。但是，量子色动力学中的手征对称性破缺却并非理想情形下的破缺，因为量子色动力学的拉氏量含有手征对称性的明显破缺项——即夸克的质量项。由于这种明显破缺项的存在，实际真空态的选取就不再是任意的了，明显破缺项的存在将会对实际真空态起到一个选择作用。这就好比一根立在桌上的筷子，如果桌子是严格水平的，它向任何一个方向倒下都是同等可能的，但如果桌子是倾斜的，它就会往倾斜度最大（梯度最大）的方向倒。用数学的语言来说（符号的含义与第七节相同），如果 $V_1(\varphi_a)$ $(a=1,2,\cdots,N)$ 表示对称性的明显破缺项，那么，它所选出的真空态将满足下列条件：

$$\Delta_a(\phi)\frac{\partial V_1}{\partial \phi_a}=0$$

这一条件被称为真空取向条件（vacuum alignment condition）。另一方面，明显破缺项的存在也破坏了戈德斯通定理成立的条件，由此导致的结果是戈德斯通粒子有可能具有非零质量，这样的粒子被称为赝戈德斯通粒子（pseudo-Goldstone particle）。真空取向条件是确定赝戈德斯通粒子质量的重要条件。赝戈德斯通粒子的出现消除了 $\pi$ 介子的非零质量与戈德斯通粒子的零质量之间的定性矛盾。但在定量上 $\pi$ 介子与赝戈德斯通粒子的质量是否吻合呢？我们现在就来看一看。

如前所述，对于量子色动力学中的手征对称性来说，对称性的明显破缺项为质量项，它可以改写成（请读者自行验证）：

---

① $\pi$ 介子的质量远小于其他强子的质量，这一点很早就引起了人们的注意。为了解释这一现象，早在量子色动力学出现之前的 1960 年，南部阳一郎就提出可能存在一种极限情形（相当于后来的手征极限），在其中 $\pi$ 介子是对称性自发破缺所产生的无质量粒子。中国物理学家周光召（1929— ）也于 1961 年提出过类似的想法。

$$V_1 = \frac{1}{2}(m_u + m_d)\bar{\psi}\psi + \frac{1}{2}(m_u - m_d)(\bar{u}u - \bar{d}d)$$

其中 $\bar{\psi}\psi=\bar{u}u+\bar{d}d$。上式的特点是：第一项只破坏手征对称性，第二项则破坏同位旋对称性。研究表明，在这些特点的基础上进一步考虑到不存在同位旋对称性的自发破缺这一限制，可以得到赝戈德斯通粒子的质量为（这一结果也可以从手征微扰理论得到）：

$$M_\pi^2 = \frac{m_u + m_d}{2F_\pi^2}\langle 0 \mid \bar{\psi}\psi \mid 0\rangle$$

其中 $F_\pi$ 是一个量纲为能量的常数，由

$$\langle 0 \mid A^{\mu a}(x) \mid \pi^b(p)\rangle = \mathrm{i}p^\mu F_\pi \delta^{ab} \mathrm{e}^{-\mathrm{i}px}$$

定义。$F_\pi$ 被称为 π 衰变常数（pion decay constant），可以由 π 介子的衰变来确定，原则上也可以从理论上计算出，其数值约为 92.4MeV①。$\langle 0|\bar{\psi}\psi|0\rangle$ 是一个量纲为能量三次方的参数，被称为手征凝聚（chiral condensation），目前人们对它的计算还比较粗略，结果大致为 $\langle 0|\bar{\psi}\psi|0\rangle \sim (270\ \mathrm{MeV})^3 n_f$，其中 $n_f$ 为参与凝聚的夸克种类，对于我们所考虑的情形 $n_f=2$（即只有 u 夸克和 d 夸克参与凝聚）②。$m_u+m_d$ 通常取为 8～9MeV。由此可以得到（请读者自己计算一下）：$m_\pi \sim 140$MeV。这几乎正好就是 π 介子的质量（$\pi^\pm$ 的质量约为 140MeV；$\pi^0$ 的质量约为 135MeV）。当然，上述估算是相当粗略的，不能因为数值上的吻合而高估它的精度。但结合了格点量子色动力学（lattice QCD）计算的大量更为细致的研究表明，这种吻合并非偶然③。

---

① 不同的文献对 $F_\pi$ 有不同的定义，彼此相差一个常数因子 2 或 $\sqrt{2}$。

② 这一结果在定性上是可以预期的，因为它大致等于量子色动力学中除夸克质量外的唯一能标 $\Lambda_{QCD}$ 的三次方。感兴趣的读者可以（定性地）思考这样一个问题：在不考虑夸克质量的情况下，量子色动力学拉氏量中唯一的参数是无量纲的耦合常数，那么像 $\Lambda_{QCD}$ 这样的能标是从何而来的？

③ 需要指出的是，对夸克质量的估计本身就在一定程度上运用了 π 介子（及其他几种介子）的质量。因此孤立地看，这里所谓的“吻合”带有循环论证的意味。但是人们对强子质量的计算是大量而系统的，涉及的粒子种类远远多于轻夸克的数目，当我们把所有这些计算综合起来看，这种“吻合”就不再是循环论证，而成为了很强的自洽性检验（consistency check）。这一点也适用于后文所述的对重子质量的计算。

现在让我们再次回到主题——质量的起源——上来。我们看到，量子色动力学计算出了作为赝戈德斯通粒子的π介子的质量。如果我们想知道π介子的质量起源，这可以算是一种回答。可惜的是，这种回答与我们在第六节中介绍的电磁自能具有相同的缺陷，那就是它正比于在理论中无法约化的外来参数：夸克质量。一旦外来参数不存在（即夸克质量为零），这一回答就会失效（因为答案也将为零）。因此量子色动力学对π介子及其他赝戈德斯通粒子质量的计算虽然很漂亮，从回答本原问题的角度看却仍不足以令人满意。

## 十三、一个 93 分的答案

但是，当我们把目光转到更复杂，同时也更具现实意义的强子——比如质子和中子（以下合称核子）——的质量时，却会看到量子色动力学的确为质量起源问题提供了一个非常精彩的回答。

计算核子或其他重子的质量是一个相当困难的低能量子色动力学问题，通常的做法是利用巨型计算机进行格点量子色动力学计算。但是，由于技术上的限制，人们在这类格点量子色动力学计算中采用的 u 夸克和 d 夸克的质量一度要比它们的实际质量高出 5 倍左右，由此得到的核子质量通常也要比实际值高出 30％以上。不过近几年，随着技术的演进，格点量子色动力学计算所采用夸克质量已逐渐降低，甚至已有一些研究者开始采用实际质量。

另一方面，与格点量子色动力学计算中夸克质量的“不可承受之重”截然相反，在我们前面提到的手征微扰理论中，夸克的质量却是越轻越好，甚至最好是零。显然，如果我们能在这两种极端之间作某种调和，借助手征微扰理论对格点量子色动力学的计算进行适当的外推，就有可能得到更接近现实世界的结果。这正是物理学家们在计算核子质量时采用的手段。这种借助手征微扰理论对格点量子色动力学计算进行外推的方法被称为手征外推（chiral extrapolation）。利用手征外推得到的核子质量为

$$m_N = m_0 - 4c_1 m_\pi^2 + O(m_\pi^3)$$

其中 $m_0 \approx 880\text{MeV}$；$c_1 \approx -1\text{GeV}^{-1}$；$m_\pi^2$ 是 π 介子的质量平方，如上节所述，正比于夸克质量。若干更高阶的项也已被计算出，这里就不细述了。将有关数据代入这一公式，我们可以得到（请读者自己计算一下）：$m_N \approx 954\text{MeV}$，它与实际的核子质量（质子约为 938MeV；中子约为 940MeV）相当接近。不仅如此，系统的计算（包括来自部分高阶项的贡献）还给出了许多其他重子的质量，比如：$m_\Sigma \approx 1192\text{MeV}$（实验值约为 $\Sigma^+$：1189MeV；$\Sigma^0$：1193MeV；$\Sigma^-$：1197MeV）；$m_\Lambda \approx 1113\text{MeV}$（实验值约为 1116MeV）；$m_\Xi \approx 1319\text{MeV}$（实验值约为 $\Xi^0$：1315MeV；$\Xi^-$：1321MeV），都与实验有不错的吻合①。这些结果表明，量子色动力学的确可以用来计算重子质量。

那么，从回答本原问题的角度看，这些计算是否令人满意呢？

从上面所引的核子质量公式中我们可以看到，上述核子质量有一个不同于赝戈德斯通粒子质量的至关重要的特点，那就是它在手征极限——即夸克质量为零——时不为零，而等于 $m_0 \approx 880\text{MeV}$。这个数值约为核子质量的 93%，它完全是由量子色动力学所描述的相互作用所确定的②。这表明，**即便不引进任何外来的夸克质量，量子色动力学仍能给出核子质量的绝大部分**。由于宇宙中可见物质的质量主要来自核子质量，因此宇宙中可见物质质量的绝大部分都可以在不引进夸克质量的情况下，由纯粹的量子色动力学加以说明。从这个意义上讲，量子色动力学为质量起源问题提供了一个独特而精彩

① 这些数值对比来自本文写作之初所参阅的文献，是大约十年前的研究结果。感兴趣的读者可以查阅一下新近文献，看是否有更好的结果。

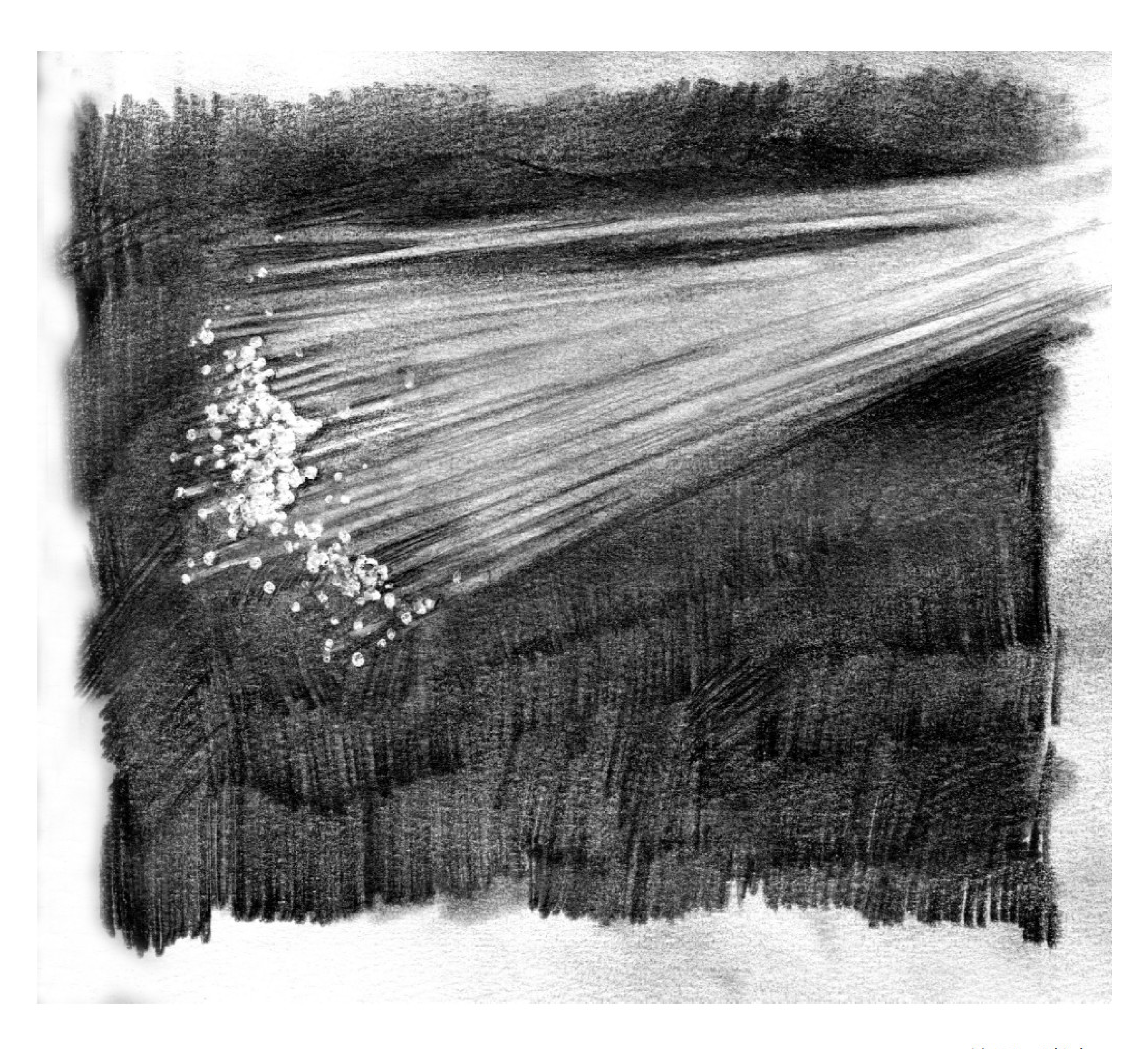
② 这个质量对应于一个由无质量的夸克和胶子组成的束缚态的质量。撇开计算上的复杂性不论，定性地讲，量子色动力学对这一质量的确定其实并不玄妙，它与量子力学对氢原子结合能的确定相类似——当然，氢原子在零质量极限下是不存在的。量子色动力学所具有的这种"质量隙"(mass gap)现象是高度非平凡的。另外，这个质量完全由相互作用所决定，在这一点上它有点类似于马赫早年的想法。只不过马赫设想的相互作用来自遥远的星体，而量子色动力学计算涉及的是微观世界的相互作用。感兴趣的读者可以思考一下：无质量的粒子为什么可以组成有质量的束缚态？

的回答。这一回答不像电弱统一理论那样带有比所要解释的质量参数还要多的可调参数，因而非常符合回答本原问题的需要。不过，由于它只能给出核子质量的93%，因此我们粗略地给它打93分。在标准模型的范围内，这是迄今所知的最佳回答。

93分虽然是一个高分，但终究不是满分。为了寻找更接近满分的答案，我们不得不重新回到标准模型中不能约化的那些质量——包括使量子色动力学丢掉7分的夸克质量——上来。那些质量究竟来自何方？究竟还能不能约化？这些问题的答案——如果有的话——就只能到标准模型之外去寻找了。

2007年1月25日写于纽约

2014年11月19日最新修订



绘画：张京

# 纤维里的光和电路中的影[①]

在一个周末的清晨，你上网查询了本市的景点信息，然后决定与家人一起参观新落成的科学博物馆；在博物馆里，你一边参观，一边用数码相机拍着像片；回家后，你用电子邮件将几张精选像片传给朋友，让他们分享你的快乐；晚上，你和家人围坐在一起，欣赏清晰的数字电视……

你也许没有意识到，在这普通的一天里，你已反复成为了2009年诺贝尔物理学奖获奖成果的受益者。

2009年10月6日，拥有英国和美国双重国籍的华裔科学家高琨(Charles K. Kao)，拥有加拿大和美国双重国籍的科学家博伊尔(Willard S. Boyle)，以及美国科学家史密斯(George E. Smith)共同荣获了2009年的诺贝尔物理学奖[②]。

---

① 本文曾发表于《科学画报》2009年第11期(上海科学技术出版社出版)。

② 由于这三位科学家的出生地及国籍丰富多彩，媒体在报道他们的获奖消息时充分发挥了灵活性。这三人在美国媒体上是三位美国科学家；在英国媒体上是一位英国科学家与两位美国科学家；在加拿大媒体上则是一位加拿大科学家与两位美国科学家。中国媒体自然也不落后，大陆媒体突出高锟的华人血统，香港媒体突出其任职香港中文大学的经历，台湾媒体则突出其“中央研究院”院士的身份。

在这三人中,高锟“因光学通信中有关光在纤维中传输的突破性贡献”(for groundbreaking achievements concerning the transmission of light in fibers for optical communication)获得全部奖金(约140万美元)的一半,博伊尔和史密斯则“因发明一种成像半导体电路——CCD传感器”(for the invention of an imaging semiconductor circuit——the CCD sensor)而分享了另一半。

在本文中,我们将对这三位科学家的工作及其意义作一个简单介绍。

## 一、光纤,信息时代的大动脉

我们先来谈谈光纤。

简单地讲,光纤是一种能引导光在其中传输的纤维。初看起来,这并不是什么深奥莫测的东西,因为光——如我们早已知道——可在一切透明介质中传输,而光纤不过是制成纤维状的透明介质。这种用介质引导光的想法早在19世纪40年代初就已出现并付诸实验(所用介质是水和玻璃),它的一种早期应用是灯光喷泉(直到今天仍在用)。由于受光纤引导的光可以随光纤而弯曲,自20世纪20年代末起,人们开始设想用光纤来制作诸如胃窥镜之类的医学仪器,那些仪器可以深入患者体内,用光纤引导的光将患处的图像传输出来。

从物理上讲,光纤利用的是一种有趣的光学现象,那就是当光从折射率较高的介质(比如玻璃)射向折射率较低的介质(比如空气)时,在特定的角度范围内,入射光会在两种介质的交界面上被完全反射,而无法进入折射率较低的介质。这种现象被称为光的全内反射(total internal reflection),如图8所示。正是它保证了光纤内的光能够被光纤所引导,而无法轻易逃逸。

事情如果仅仅是这样,就没诺贝尔奖什么事了。人们在实际制作光纤时很快就发现,虽有全内反射在光纤的边界上把关,光纤中的光仍会迅速损耗。在20世纪60年代初,光在最好的光纤中,也只能传播区区20米就只剩下了1%左右。这使得光纤的应用只能局限于像医学仪器那样的短距离之内。

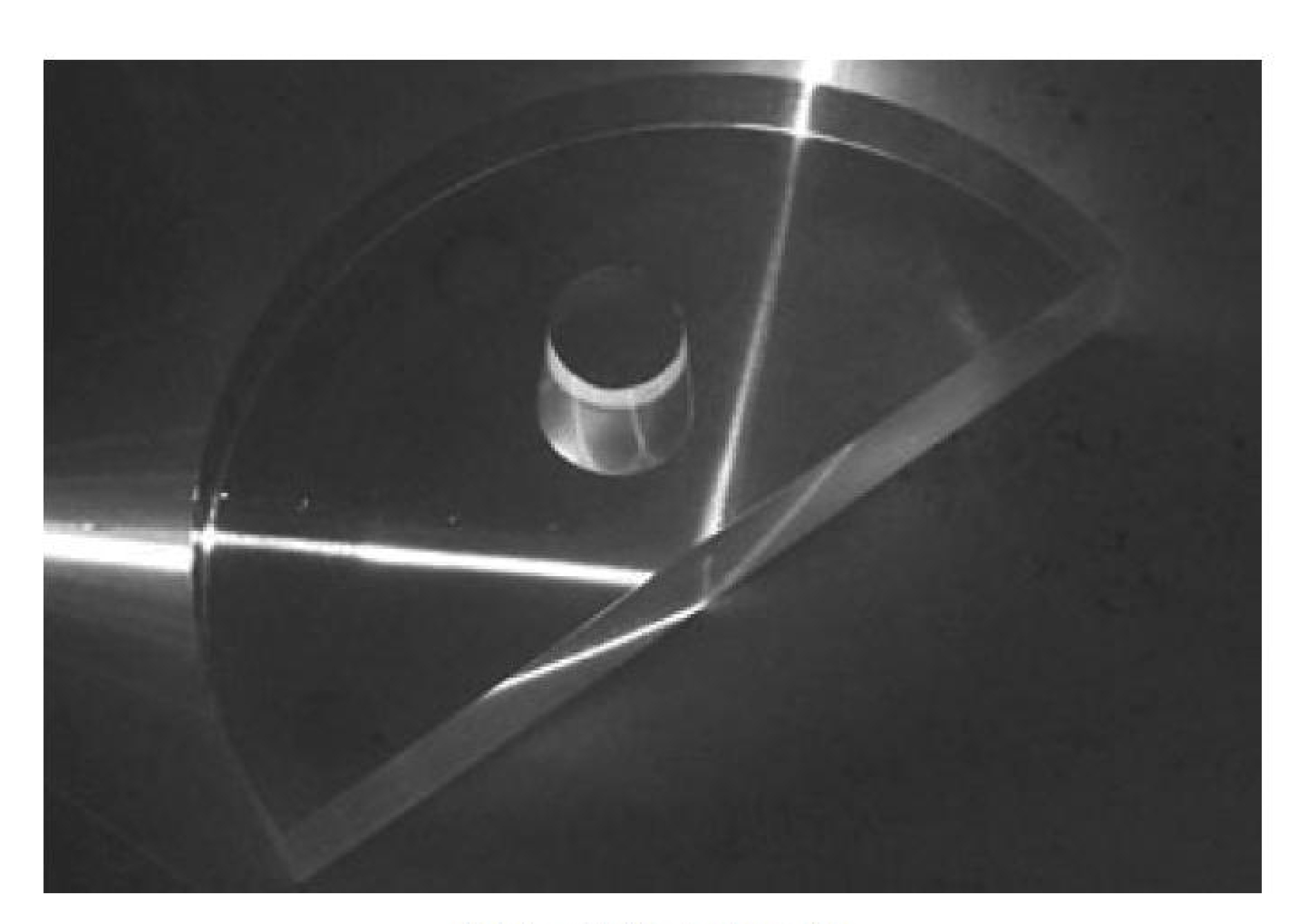


图 8　光的全内反射

那么，光纤中光的快速损耗究竟是什么造成的呢？人们提出了一些可能的原因，比如光纤的弯曲，或光纤材料（比如二氧化硅）的晶体结构缺陷等。但是，任何实际应用中的光纤都不可能不弯曲，任何常温下的晶体结构也都不可能无缺陷。因此，若原因果真在这些方面，那光的快速损耗基本上就是“绝症”了。幸运的是，就在这光纤应用的整体前景面临极大挑战的时候，英国标准电信实验室（Standard Telecommunications Laboratories）的高锟与霍克汉姆（George Hockham）经研究发现①，光的快速损耗并非上述原因所致，而主要是由于光纤中杂质——尤其是铁离子——对光的吸收与散射。他们这项研究为光纤时代的降临开启了大门②，因为既然罪魁祸首是杂质，我们要做的就只

① 霍克汉姆于 1969 年获得电子工程学博士学位，一生获得过 16 项专利。高锟曾在 2004 年的一次访谈中提到，霍克汉姆从事的是理论研究。高锟成为当年那项研究的唯一获奖者，有可能是因为霍克汉姆当时还只是一位研究生。诺贝尔奖有过忽略研究生的先例，比如英国天文学家休伊什（Anthony Hewish）因脉冲星的发现而获得了 1974 年的诺贝尔物理学奖，他的学生贝尔（Jocelyn Bell Burnell）虽然是实际上的发现者，却没有获奖。当然，高锟在那篇论文发表之后又与其他人合作，对其他材料、其他波长的光纤应用进行了研究，为工业界指出了更具体的努力方向，这也很可能是他成为那项研究的唯一获奖者的原因。

② 高锟被一些媒体称为“光纤之父”，不过“光纤之父”之名在此次诺贝尔物理学奖公布之前，通常是指美籍印度裔科学家卡潘尼（Narinder Kapany），他在 20 世纪 50 年代做过很多光纤方面的工作。另外要提到的是，与高锟的研究同年，德国科学家伯尔纳（M. Boerner）也提出了类似的观点，并在德、英、美等国获得了专利，不过此人不久后就去世了。

是对光纤材料进行提纯，而这是没有任何原则性困难的。

高锟等人的工作发表于1966年。4年之后，即1970年，美国玻璃制造商康宁公司就通过材料提纯，将原先20米的传输距离提升到了1000米①。此后，就像所有技术领域的发展一样，这一纪录被一再刷新。自1975年起，英、美、日等国先后迈出了实用光纤通信的步伐。1988年，第一条跨大西洋的光纤电缆安装成功。现代的互联网、有线电视、电话通信等更是处处离不开光纤(图9)。可以毫不夸张地说，光纤已成为信息时代的大动脉。与传统的无线电通信相比，光纤所能传输的信息量要大得多，而且光纤所用之材料不仅蕴藏丰富，而且强度很高，具有得天独厚的应用优势。据估计，人们迄今铺设的光纤网络已达10亿千米左右，足可在地球与月亮之间绕一千多个来回。

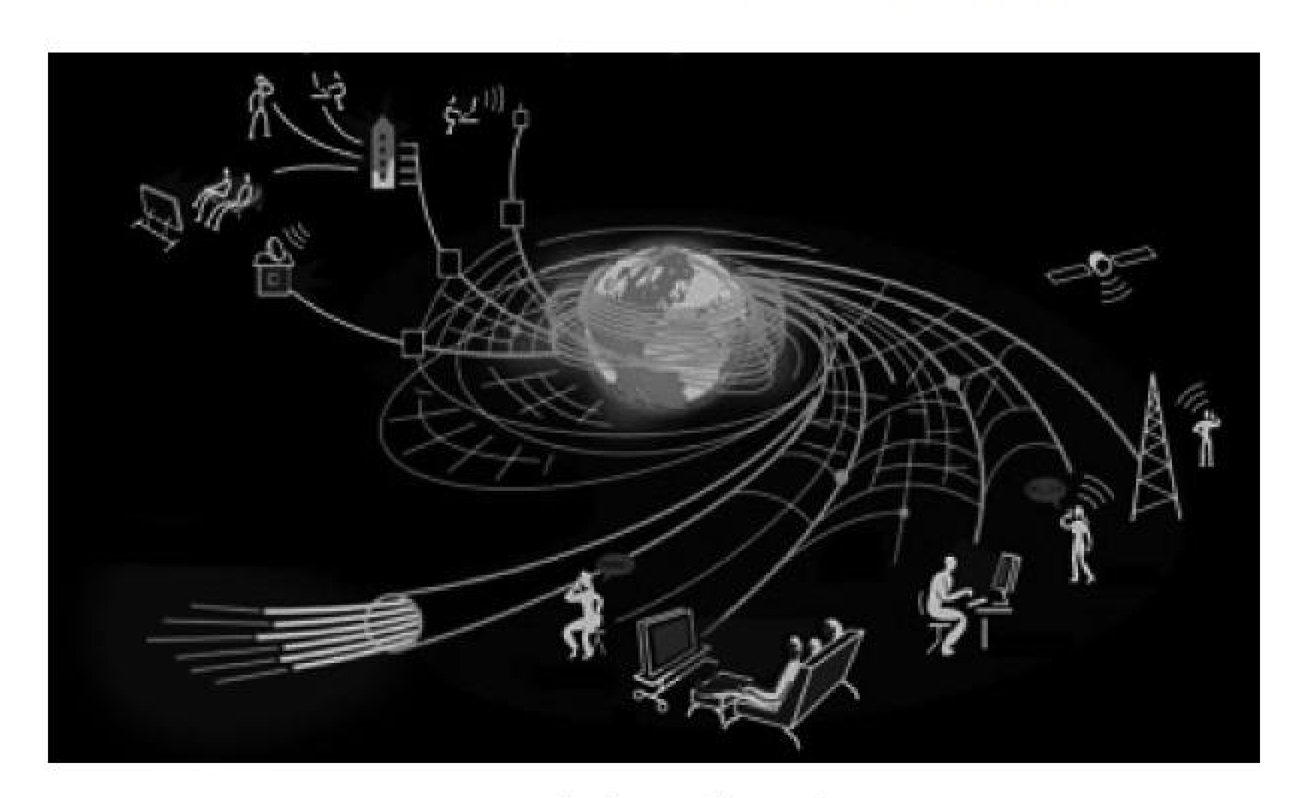


图9　光纤网络示意图

在光纤所传输的信息里，有很大一部分是数码影像，这些影像的由来将我们引向了今年诺贝尔物理学奖的第二项获奖工作：CCD。

## 二、CCD，数码摄影的电子眼

CCD是电荷耦合器件(charge-coupled device)的英文缩写。这种器件原本是作为一种电子内存而研发的。1969年秋天，美国贝尔实验室的博伊尔

① 用技术性的术语来说，康宁公司将光纤的损耗系数由每千米1000分贝减少为了17分贝。

(Willard S. Boyle)和史密斯(George E. Smith)从事的就是这种研发工作。但 CCD 的真实用途几乎立刻就转变为了感光器件。

CCD 的感光原理是建立在一种被称为光电效应(photoelectric effect)的现象之上的。这种现象曾被电磁波的发现者，德国物理学家赫兹(Heinrich Hertz)观察到——因此有时也被称为赫兹效应(Hertz effect)，后来又经过了实验物理学家勒纳(Philipp Lenard)的研究，并由爱因斯坦利用当时还很新颖的光量子理论作出了理论解释(勒纳德与爱因斯坦因此分别获得了 1905 和 1921 年的诺贝尔物理学奖)。按照光电效应，适当频率的光照射到某些物质上时，会从物质中打出电子，其数目与光强成正比。

利用这一效应，博伊尔和史密斯将感光材料制成了一个由很多小单元组成的阵列①，当光照射到阵列上时，会在每个小单元上打出一些电子。这些电子的数目分布很好地记录了入射光的强度分布。为了保存这些电子，博伊尔和史密斯让每个感光小单元都配有一个微小的电容。在感光过程结束后，这些小电容里的电子通过巧妙设计的电路逐排传递出去，并转变成为数字信号。这就是 CCD 的工作原理，而由那些数字信号组成的就是所谓的数码影像。由于 CCD 所用的将电子逐排传递出去的方式很像早年消防队员人工传递水桶的情形，因此这种器件也被称为“组桶式”器件(bucket brigade device)，如图 10 所示。

萌生 CCD 设想后的第二年，博伊尔和史密斯就将它用到了摄像机上；1972 年，一家美国公司率先制造出了具有 10 000(100×100)个感光单元的 CCD 传感器；1974 年，第一张 CCD 天文相片问世；1975 年，CCD 摄像机达到了可用于电视转播的水准；1979 年，CCD 被首次安装到了天文望远镜上……CCD 的发展走上了快车道。近年来，在 CCD 的冲击及其他因素的影响下，世界最大的胶卷生产商柯达公司(Eastman Kodak Company)陆续停止了普通胶

① 感光材料的选取标准是在所需的频率范围——比如可见光区——内具有显著的光电效应。



图 10　CCD"组桶式"传输电子的比喻图

片及胶片相机的生产。从某种意义上讲,这意味着一个时代——光学摄影时代——的终结。当然,它同时也是一个新时代——数码影像时代——日益成熟的标志。

那么,年轻的 CCD 与历史悠久的普通胶片相比究竟有什么优点呢？主要的优点有两个：一个是敏感度高,CCD 能对 90%左右的入射光子产生反应,也就是说,100 个入射光子约有 90 个能在 CCD 的感光材料上产生电子,从而得到记录。而普通胶片及肉眼只能记录其中 1～2 个(高质量的胶片也只能记录 10 个左右)。另一个是适用范围广,CCD 可用于从红外到 X 射线的各种波段。而普通胶片的适用范围却很狭窄,早期的普通胶片甚至无法有效地涵盖可见光区内的红光,从而使得像褐矮星、红移值较高的类星体之类偏于长波的天体的发现大大延后。此外,普通胶片需要冲印,这对日常使用来说虽只是小麻烦,但对行星探测器来说可就要了命了,因为行星探测器大都是一去不复返的,不可能将胶片带回地球冲印。而 CCD 的数码信息却可以通过电波传回地球。我们今天看到的那些美轮美奂的行星图片,或哈勃太空望远镜(Hubble space telescope)拍摄的遥远星云都是因为有了 CCD 这只电子眼才成为了可能。对于观测天文学来说,CCD 是一项能媲美望远镜与光谱仪的伟大发明。

光纤通信与 CCD 都是技术成就，但它们对于科学研究同样是必不可少的。今天的科学家们每天都在通过光纤大动脉交流着研究信息；翱翔在外层空间的太空望远镜每天都在用 CCD 电子眼窥视着这个让人着迷的宇宙。从这个意义上讲，获得今年诺贝尔物理学奖的虽是技术领域的工作，却对科学的发展有着意义深远的促进。

## 附录：获奖者小档案



高锟　　博伊尔　　史密斯

- 高锟(Charles K. Kao)：拥有英国和美国双重国籍的华裔科学家，1933 年 11 月 4 日出生于中国上海，1965 年获伦敦帝国大学(Imperial College London)电子工程学博士学位。高锟曾在英国标准电信实验室(Standard Telecommunications Laboratories)、香港中文大学等处任职，1996 年退休，目前居住在美国。
- 博伊尔(Willard S. Boyle)：拥有加拿大和美国双重国籍的科学家，1924 年 8 月 19 日出生于加拿大艾姆赫斯特(Amherst)，1950 年获加拿大麦吉尔大学(McGill University)物理学博士学位。博伊尔自 1953 年起在美国贝尔实验室(Bell Labs)任职，期间曾于 20 世纪 60 年代参与阿波罗登月计划，1979 年退休，目前居住在加拿大。

- 史密斯(George E. Smith):美国科学家,1930 年 5 月 10 日出生于美国怀特普莱恩斯(White Plains),1959 年获美国芝加哥大学物理学博士学位。史密斯自 1959 年起在美国贝尔实验室(Bell Labs)任职,期间获得过几十项技术专利,1986 年退休,目前居住在美国。

2009 年 10 月 11 日写于纽约

# 石墨烯——从象牙塔到未来世界[①]

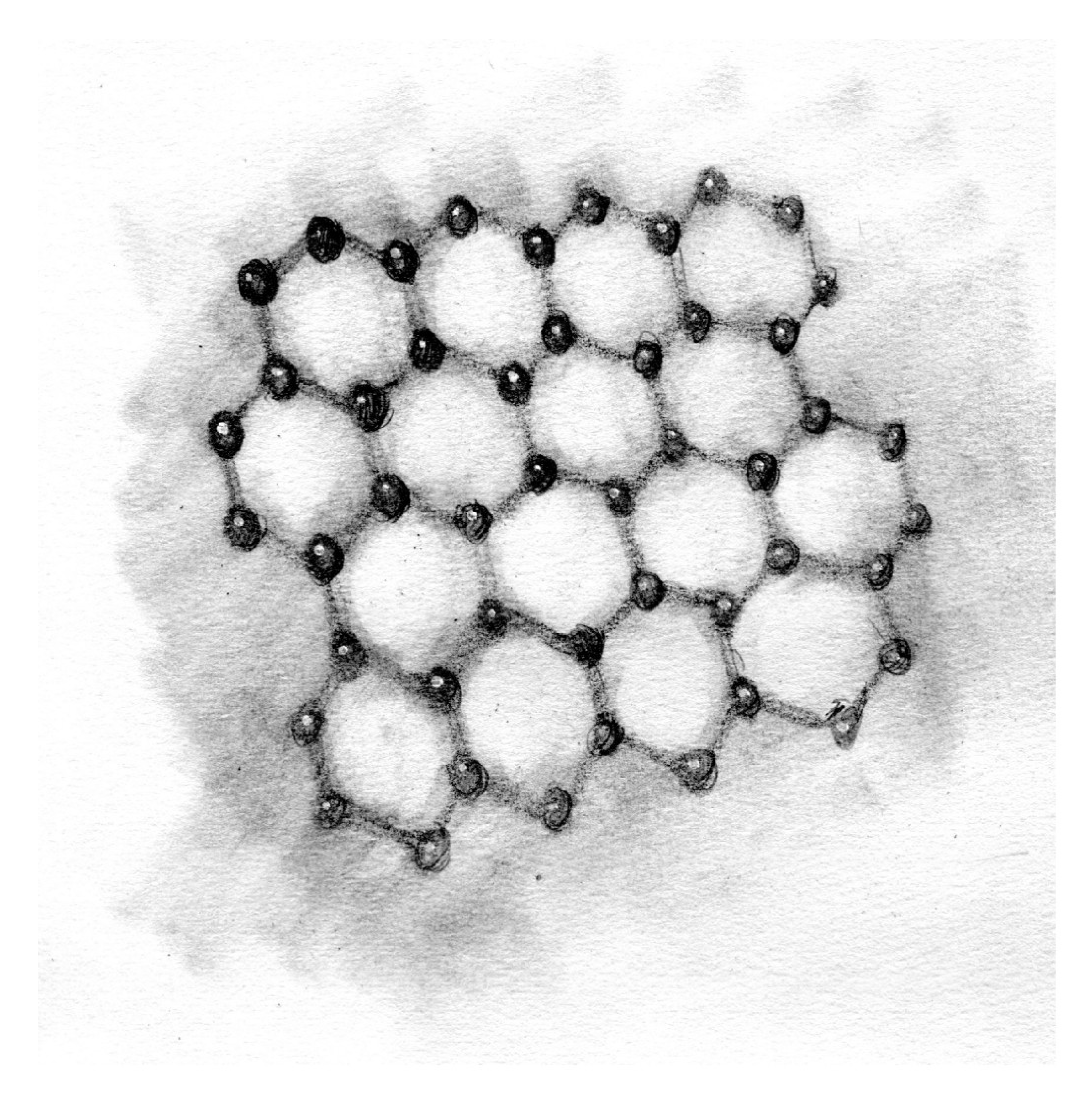
2010 年 10 月 5 日，瑞典皇家科学院（The Royal Swedish Academy of Sciences）宣布了 2010 年诺贝尔物理学奖的得主。荷兰籍俄裔物理学家盖姆（Andre Geim）和拥有俄罗斯及英国双重国籍的物理学家诺沃肖洛夫（Konstantin Novoselov）由于“对二维材料石墨烯的突破性实验”（for groundbreaking experiments regarding the two-dimensional material graphene）而共同荣获了这一奖项。

在本文中，我们将对这两位物理学家的获奖成果及其意义作一个简单介绍。

## 一、来自象牙塔的新材料

我们先来说明一下什么是石墨烯。这个名称中的“石墨”（graphite）二字我们大都不陌生，因为铅笔的笔芯就是由它和黏土混合而成的。从元素的角

① 本文曾发表于《科学画报》2010 年第 11 期（上海科学技术出版社出版）。



绘画：张京

度讲，石墨是由碳元素组成的。在电子显微镜下，我们可以发现石墨的结构是层状的，每一层的碳原子都排列成紧密的蜂窝状六边形网格，层与层之间的距离则比较大，形成松散的堆砌①(图 11)。铅笔之所以在纸上轻轻一划就会留下痕迹，正是这种松散堆砌的结果。那么石墨烯(graphene)又是什么呢？它就是单层的石墨。

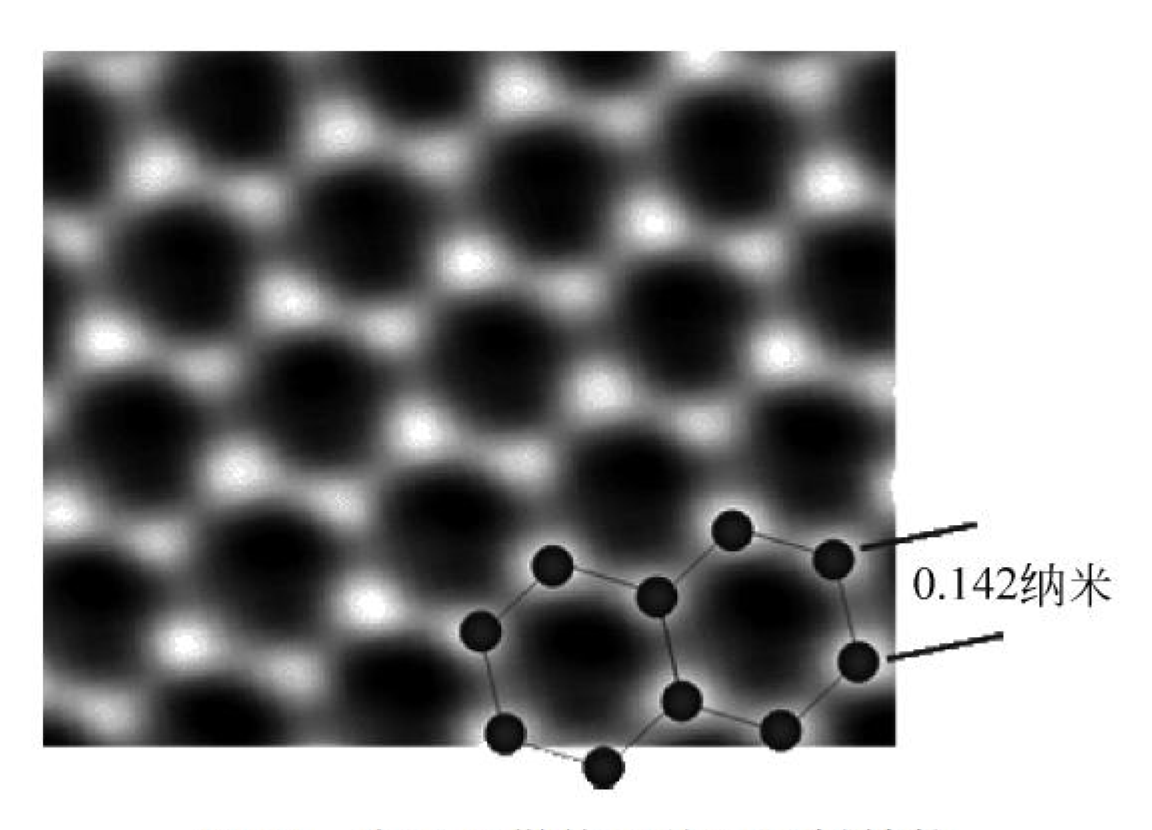



图 11　电子显微镜下的石墨烯结构

石墨烯这个名称是从 1987 年开始使用的，但在那之前，就已经有人对这种单层原子组成的二维结构产生了兴趣，因为这种结构比现实世界里的三维结构来得简单，很适合当作例题收录在教科书里②。通过这种象牙塔式的兴趣，人们开始对石墨烯的性质有了一些理论上的了解。这种了解，加上技术领域对新材料的需求日益旺盛，使人们对石墨烯产生了更现实的兴趣，试图将它由单纯的象牙塔物质“提拔”为真实材料。

初看起来，这种“提拔”似乎不会太困难。事实上，当我们用铅笔在纸上轻轻划过时，划痕中就可能会出现单层的石墨——即石墨烯。但问题是，铅笔的划痕从微观角度讲实在是太大了，在那里搜寻石墨烯简直就像是在整个喜马拉雅山脉中搜寻一片薄冰，即便找到也只能算是瞎猫碰上了死耗子。而科学家们

---

①　石墨每一层上的碳原子间距约为 0.142 纳米，层与层的间距则为 0.335 纳米，后者是依靠微弱的范德瓦耳斯力(van der Vaals force)结合起来的，因而是松散的堆砌。

②　当然，这里所谓的“二维”不是几何上的二维，而仅仅是指垂直方向上的物理自由度可以忽略的情形。

需要的是系统的方法，是可以复制的成功，这却是相当困难的。直到 21 世纪初，人们所达到的最好业绩——即最薄的石墨片——也只能薄到几十层原子的水平。

更糟糕的是，有迹象表明，像石墨烯那样的二维材料有可能是注定只能存在于象牙塔里的。因为早在 20 世纪 30 年代，著名俄国物理学家朗道（Lev Landau）等人就已证明，二维材料的热运动涨落会破坏自身的结构。实验上制备石墨烯的种种失败尝试似乎也在佐证着这一结论，比如石墨层越薄，就越容易卷曲成球状或柱状，而无法维持平面结构①。因此，制备石墨烯曾被很多人认为是注定无法成功的。

但以盖姆为核心的实验组却不信这个邪，决意尝试这一看似不可能的任务。这种尝试对他们来说，乃是一贯作风的延续。因为在盖姆的实验组里，对各种有趣、甚至有趣得近乎荒谬的事情的尝试已经达到了制度化的程度，他们每星期都几乎固定地拿出十分之一的时间来做一种所谓的“星期五之夜实验”（Friday evening experiment），专门尝试各种稀奇古怪的事情②。制备石墨烯的工作也是从一个“星期五之夜实验”开始的。经过一些失败的尝试后，他们采用了所谓的“透明胶大法”（Scotch tape technique），即用透明胶粘住石墨层的两个面，然后撕开，使之分为两片。通过不断重复这一“大法”，并辅以其他手段，他们最终制备出了石墨烯③。

---

① 不过那种球状或柱状的结构对于石墨烯的制备来说虽是“麻烦制造者”，本身却都是绝顶的好东西：前者是所谓的富勒烯（fullerene），它的发现者获得了 1996 年的诺贝尔化学奖；后者则是大名鼎鼎的纳米管（nanotube），也是一种令人着迷的新材料。

② 盖姆曾经因为在这种“星期五之夜实验”中进行过“磁悬浮青蛙”实验，而获得了 2000 年的搞笑诺贝尔物理学奖（Ig Nobel Prize in Physics）。他是迄今唯一一位同时获得过搞笑诺贝尔奖和诺贝尔奖的人。

③ 有读者可能会问：既然朗道曾经证明过二维材料的涨落会破坏物质结构，怎么还可能制备出石墨烯呢？答案是，朗道的证明是针对大面积（理论上是无穷大）的体系的，而人们最初制备的石墨烯只有几平方微米。另一方面，朗道的证明考虑的是严格的平面，而真实的石墨烯会在三维空间里波动，从而耗散掉一部分涨落能量。因此石墨烯的出现虽然出人意料，却不是不可理解的。

盖姆和诺沃肖洛夫获奖后，许多媒体推出了渲染性的标题，比如《物理学家用透明胶和铅笔赢得诺贝尔奖》。这种标题容易给人一个错觉，以为那是一项轻而易举的工作。事实上，盖姆实验组制备石墨烯的过程并不轻松，前后持续了一年多的时间，制备出的石墨烯则只有几平方微米，要用高倍显微镜才能观测。而且由于石墨烯是高度透明的，在观测及制备过程中还有一个如何分辨的问题。盖姆实验组解决这一问题的方法，是巧妙地利用了石墨烯在厚度300 纳米的二氧化硅晶片衬底上产生的光线干涉效应。这一点是他们胜过其他研究组的关键所在。但即便如此，他们当时选用的衬底如果不是二氧化硅而是其他晶片，或者晶片的厚度不是 300 纳米，而是略大或略小，就都有可能无法分辨石墨烯。而他们当时之所以选用了恰到好处的衬底，据诺沃肖洛夫回忆乃是纯属偶然。因此，盖姆实验组的成功背后既有长时间的努力和巧妙的构思，也有运气的成分①。当然，既然想到了正确的方法，发现合适的衬底应该是迟早的事情，从这点上讲，他们的成就并非偶然。

那么，这种辛辛苦苦制备出来的二维材料在我们这个三维世界里究竟有什么用处呢？在现实的用处出现之前，它在理论上的用处就已经吸引了科学家们的兴趣。物理学家们早在 1956 年就发现，托二维世界的福，石墨烯中的电子运动具有很奇特的性质，即电子的质量仿佛是不存在的②。这种性质使石墨烯成为了一种罕见的可用于研究所谓相对论量子力学的凝聚态物质——因为无质量的粒子必须以光速运动，从而必须用相对论量子力学来描述。而更奇妙的是，那种相对论量子力学中的“光速”并不是真空中的光速，而只有后者的 1/300。很多科学爱好者也许读过俄国物理学家伽莫夫（George

① 制备石墨烯——尤其是大样品——的难度还可以从另一个角度来印证，那就是石墨烯的价格。直到 2008 年 4 月，石墨烯的价格依然高到令人瞠目的每平方厘米一亿美元，堪称史上最贵的材料。不过最近两年，人们制备石墨烯的能力已突飞猛进，最大样品的线度已超过 70 厘米，价格也已暴跌（因此千万不要囤积石墨烯，它很重要，但绝不可能使你发财）。

② 确切地说，那并非电子，而是电子与石墨烯晶格相互作用所产生的准粒子（quasi-particle），是石墨烯的低能激发态。

Gamow)所写的科普作品《物理世界奇遇记》(*Mr. Tompkins in Paperback*),在那部作品中伽莫夫设想过一个光速很缓慢的世界。从某种意义上讲,石墨烯就是那样一个世界,它所具有的奇妙性质为理论物理学家们提供了一片研究相对论量子力学的新天地,使他们不仅可以把一些原先要用巨型加速器来研究的问题搬到自己的小型实验室里,而且还可以研究一些用巨型加速器都未曾有机会透彻研究的东西,比如所谓的克莱因佯谬(Klein's paradox)或相对论量子力学特有的所谓"颤振"(zitterbewegung)效应,甚至还可以研究弯曲空间里的相对论量子力学——因为在石墨烯这个舞台上,弯曲空间不过就是弯曲的石墨烯而已。这些理论研究不仅题材新颖,而且还特别便于观测,因为石墨烯是二维的,所有现象都出现在表面上,不会像三维材料中的现象那样有可能跑到物质内部去。

除了成为研究相对论量子力学的新天地外,石墨烯还具有所谓的量子霍耳效应(quantum Hall effect),这种本身就是诺贝尔奖量级的重要效应以往是要在极低温下才能显现的,石墨烯却能将它带到室温下。诺沃肖洛夫在接受媒体采访时曾经表示,要让物理学家们改变自己的研究方向,必须用比他们所研究的有趣十倍的东西来引诱。石墨烯对很多理论物理学家来说看来就具有那样的魅力,因而吸引了众多的追随者。

## 二、通往未来世界的金桥

但石墨烯最吸引人的地方还在于它在现实世界里的可能应用。由于石墨烯的结构极为紧密和严整,哪怕在室温下都几乎没有任何缺陷,最大限度地发挥了众原子"集体的力量",这使它不仅有比同等线度的钢铁还高两个数量级的强度,而且还有普通刚性材料难以企及的韧性,可以拉伸20%而不断裂。显示这种性质的流传最广的图片,是一幅猫躺在石墨烯制成的吊床上休息的想象图。这种由单层原子制成的吊床居然可以承受宏观物体的重量,无疑是令人惊叹的。那幅图片不够确切的地方,是没能显示出石墨烯的超薄特性。

由于石墨烯的透光率高达97.7%①，厚度却只有单层原子，因此如果真有那样的吊床，它不仅对于肉眼，甚至对于很多仪器都会是不可见的，我们看到的将是一只悬停在半空中的猫，就像《爱丽丝漫游奇境记》（*Alice's Adventure in Wonderland*）里那只柴郡猫（Cheshire cat）的笑容一样。

石墨烯如果只用来制作吊床，那显然是大材小用了。它更重要的可能应用是制成超薄、超轻、超强的材料，用于飞机、火箭、防弹衣等对材料性质要求极高的产品中。而它最能扣动人们想象之弦的可能应用，则是所谓的太空电梯。这种早在1895年就由火箭理论的先驱者、俄国科学家齐奥尔科夫斯基（Konstantin Tsiolkovsky）提出过的迷人设想，一直面临着一个致命问题，那就是找不到具有足够强度的材料来支撑线度达几万千米的巨型结构。石墨烯的出现使很多人重新燃起了希望。

除上述可能应用外，石墨烯的另一类可能应用则倚仗于它的电子运动性质。如我们在前面所述，石墨烯中的电子运动具有很奇特的性质，比如电子的质量仿佛是不存在的，而运动速度是所谓的"光速"。这些特性，加上石墨烯结构在常温下的高度完美性，使得电子的传输及对外场的反应都超级迅速，几乎达到了人们梦寐以求的境界。体现到物理性质上，这使得石墨烯具有超常的导电性和导热性。这种性能既体现在纯净的石墨烯中，也可以部分地体现在含有石墨烯的复合材料中。而且更重要的是，石墨烯还可以用来制作晶体管，由于石墨烯结构的高度稳定性，这种晶体管在接近单个原子的线度上依然能稳定地工作。相比之下，目前勇挑大梁的以硅为材料的晶体管在10纳米（相当于几十层原子）左右的尺度上就会失去稳定性；而石墨烯中电子对外场的反应速度超快这一特点，又使得由它制成的晶体管可以达到极高的工作频率。事实上，IBM公司在2010年2月就已宣布将石墨烯晶体管的工作频率提高

① 石墨烯的这个透光率（对应于吸收率2.3%）是一个漂亮的理论结果，精确公式为$(1+\pi\alpha/2)^{-2}$，其中$\alpha$（$\approx 1/137$）是所谓的精细结构常数。很多媒体引用的是这一公式的近似式：$1-\pi\alpha$。

到了 1000 亿赫兹，超过了同等线度的硅晶体管①。很多人相信，石墨烯将会成为硅的接班人，引领技术领域一个新的微缩时代的来临。

石墨烯的可能应用还有很多，比如它除了具有超高的强度和韧性外，还有不透水、不透气，以及抵御强酸、强碱的能力，这使它有可能成为制作保护膜的理想材料。而石墨烯既能导电又高度透明的特点，则使它有可能在制作液晶显示屏、触摸显示屏、太阳能电池板等领域大显身手。此外，用石墨烯制作的能快速充电的电池、容量超高的电容、能检测单个污染物分子的污染探测器、能用于量子计算机的特殊元件等，也都在构想或研制之中。

石墨烯从制备到获奖只用了短短六年的时间，与动辄要回溯几十年去“考古”的前几年的获奖成果相比，是非常快的。但在这六年里，由它开启的研究领域呈现了井喷的势头，几乎每个月都有新兴的研究方向被开辟出来。也许在不太遥远的将来，我们会开着由石墨烯电池驱动的车子去上班，在由石墨烯太阳能板提供能源的办公室里，用“内含石墨烯”（Graphene Inside——取代 Intel Inside）的计算机从事工作。在假日里——如果有闲钱的话——我们也许还可以乘坐用石墨烯材料建造的太空电梯去地球同步轨道欣赏地月同辉的奇景。这一切奇思妙想都得益于六年前的那项工作。在有关未来世界的构想中，很少有一种材料能像石墨烯那样大范围、跨领域地激发人们的想象力，并使人们因为看到实实在在的希望而有可能投入实实在在的努力。从这个意义上讲，它仿佛一座通往未来世界的金桥。

## 附录：获奖者小档案

- 盖姆（Andre Geim）：荷兰籍俄裔物理学家，1958 年 10 月 1 日出生于俄国城市索契（Sochi），1987 年获俄国科学院固体物理研究所博士学

---

① IBM 所宣称的 1000 亿赫兹其实是“适度浮夸”的结果，实际试验中所达到的频率约为 300 亿赫兹。


盖姆


诺沃肖洛夫

位。自 1990 年起，盖姆先后在英国诺丁汉大学（University of Nottingham）、丹麦哥本哈根大学（University of Copenhagen）、英国巴斯大学（University of Bath）、荷兰内梅亨大学（Radboud University Nijmegen）等地工作过。2001 年，盖姆成为英国曼彻斯特大学（University of Manchester）物理学教授，并于 2002 年起担任曼彻斯特介观科学及纳米技术中心（Manchester Centre for Mesoscience and Nanotechnology）主任。

- 诺沃肖洛夫（Konstantin Novoselov）：拥有俄罗斯及英国双重国籍的物理学家，1974 年 8 月 23 日出生于俄国城市尼茨塔吉尔（Nizhny Tagil），2004 年获荷兰内梅亨大学博士学位。诺沃肖洛夫是盖姆的学生及长期合作者，自 2001 年起，与盖姆一起在英国曼彻斯特大学工作。诺沃肖洛夫是自 1973 年以来最年轻的诺贝尔物理学奖得主。

2010 年 10 月 11 日写于纽约

# 囚禁的量子,开放的应用①

2012 年 10 月 9 日,一位 68 岁的法国老人与妻子在街头散步,当他们路过一条街边的长椅时,电话忽然响起,老人被告知获得了诺贝尔物理学奖。同样被"搅扰"的还有大西洋彼岸的一位也是 68 岁的美国老人,电话响起时他还在睡梦中,但无论什么梦也没有电话里的消息更美:他也获得了诺贝尔物理学奖。

这两位天各一方,但恰巧同岁的老人分别是法国物理学家阿罗什(Serge Haroche)和美国物理学家维因兰德(David Wineland),之所以获奖,是因为他们实现了"使得对单个量子体系的测量与操控成为可能的突破性实验方法"(for ground-breaking experimental methods that enable measuring and manipulation of individual quantum systems)。他们将共同分享崇高的荣誉,以及虽因金融危机而缩水,但数量依然可观的 800 万瑞典克朗(约合 110 万美元)的奖金②。

在本文中,我们将对这两位物理学家的工作及其意义作一个简单介绍。

---

① 本文曾发表于《科学画报》2012 年第 11 期(上海科学技术出版社出版)。

② 在过去若干年里,每个奖项的奖金为 1 000 万瑞典克朗。

## 一、小有小的麻烦

美国物理学家费恩曼曾以一个有趣的问题作为《费恩曼物理学讲义》(*The Feynman Lectures on Physics*)的开篇，那就是：假如因为某种灾变，在所有科学知识中只有一句话能传之于后代，什么话能用最少的文字包含最多的信息？费恩曼认为，那应该是所谓的"原子假设"，即所有物质都是由原子组成的[①]。不过，这句话包含的信息虽多，要想破译却并不容易。事实上，早在两千多年前的古希腊就有先贤猜测过物质是由原子组成的("原子"一词的英文 atom 就来自希腊文 τομο ς，含义为"不可分割的")，但直到 18 世纪才开始有了现代意义下的原子理论，而原子的真正奥秘，则直到 20 世纪才开始揭晓。

为什么呢？因为原子实在太小了，既看不见，也摸不着。

如今我们知道，原子并非是"不可分割的"，它由更基本的粒子所组成，并且与那些粒子一样，遵守一种被称为量子力学(quantum mechanics)的奇妙规律。这种规律与我们习以为常的宏观世界的规律完全不同，在发现之初曾带给物理学家们极大的震动。直到很多年后，当那种规律逐渐褪去新鲜的外衣，甚至已变成物理系学生的常识时，想在最直接的意义上体验它们仍是极为困难的事情。

为什么呢？依然是因为原子实在太小了，既看不见，也摸不着。

由于这一原因，物理学家们对原子——或者更一般的，对量子体系——的很多观测都不是针对单个原子(或量子体系)的。比如他们观测的原子光谱乃是由很多原子共同发射的。而在有条件观测单个原子(或量子体系)的实验中，由于观测对象太小，往往观测一结束，观测对象本身也就"人间蒸发"或"香消玉殒"了，比如用云室或气泡室(这两者的发明者分别获得了 1927 年和

---

① 这里我们稍稍偷了点懒，费恩曼想要传给后代的话还包括了原子处于永恒的运动之中，以及它们太过靠近时彼此排斥，稍稍远离时彼此吸引这几点。

1960 年的诺贝尔物理学奖）观测粒子，或用照相设备观测光子就都是如此。

那么，有没有什么办法，能够观测甚至操控单个量子体系，同时还让它继续存在（从而还可以继续观测或操控）呢？维因兰德和阿罗什——在他们各自同事的鼎力合作下——所解决的正是这个问题。他们凭借高超的实验技巧，将单个量子体系囚禁起来，然后用细微而巧妙的“探针”去观测甚至操控它，从而完成了近乎“不可能任务”（mission impossible）的壮举，为上述问题提供了肯定答案（图 12）。

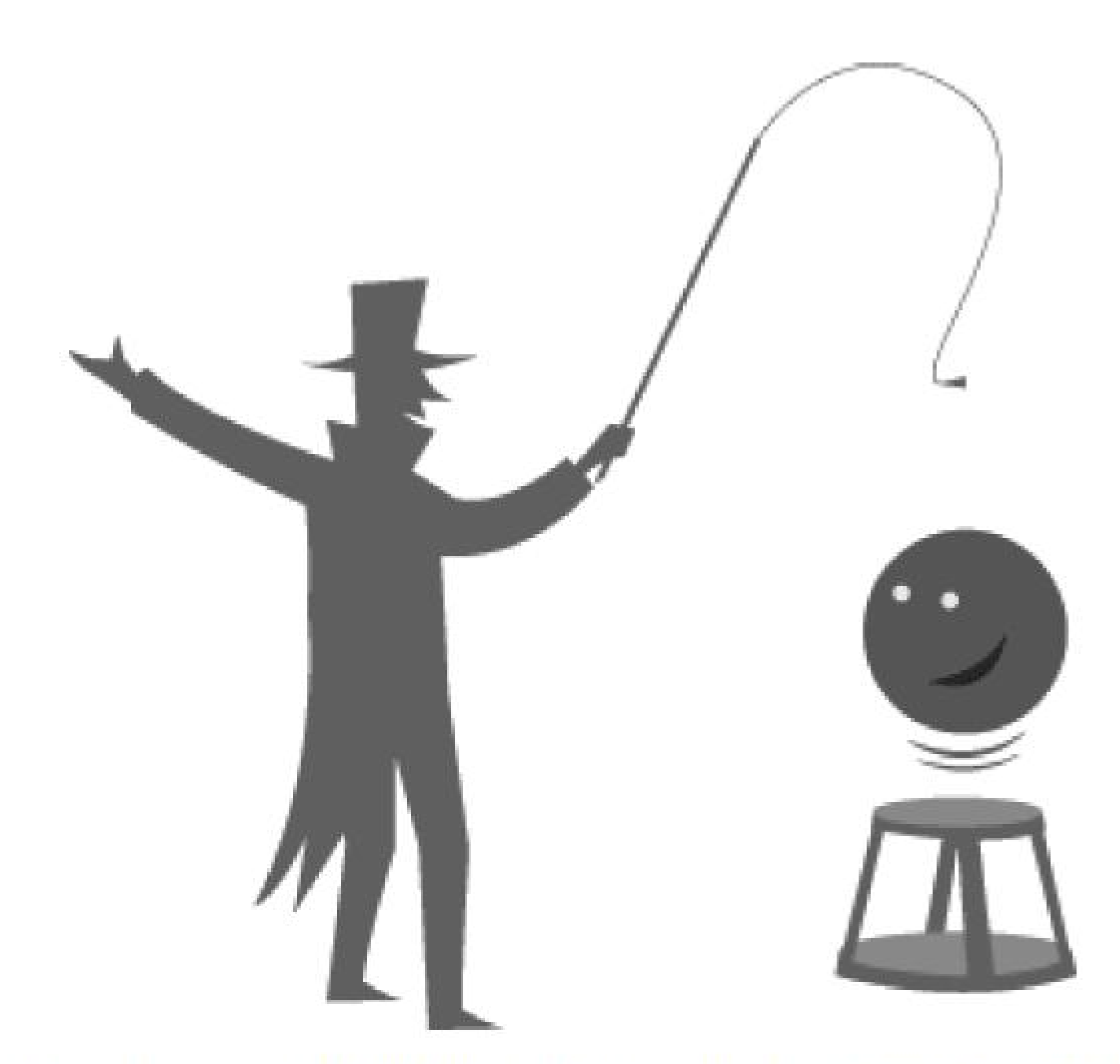


图 12　维因兰德和阿罗什完成了近乎“不可能任务”的壮举

下面我们就对他们的方法做一个简单介绍。

## 二、囚禁的量子

维因兰德采用的方法是将单个的离子（离子是失去或得到若干电子——从而带电——的原子），比如铍离子 $Be^+$（它是失去一个电子的铍原子），利用其带电的特征，囚禁在用电磁场组成的“牢笼”中，然后以光子作“探针”去探测和操控它。这话说起来简单，实现起来却极不容易，单是那“牢笼”——它的“学名”叫做离子阱（ion trap）——本身就已是一个诺贝尔奖级别的成就（它的

实现者获得了1989年的诺贝尔物理学奖)①。为了确保被囚禁的是单个(或少数几个)离子,还需要辅以超高真空(以便排除其他粒子的干扰)和超低温(以便排除热运动的干扰)等技术。其中后者采用的乃是维因兰德与同事亲自参与研发的绝活:边带冷却技术(sideband cooling)②。当这些极不简单的配置完成之后,维因兰德又通过激光脉冲(光子),将被囚禁离子的内部状态(即电子能态)叠加起来。这种状态叠加是量子力学有别于经典物理的奇妙特征,科普读物中常见的诸如“粒子既在这里,又在那里”,“猫既是死的,又是活的”,等等吸引眼球的表述都源自于此。但维因兰德能做到的还不止这些,通过对激光脉冲的巧妙选择,他还可以对状态叠加的方式进行操控,比如将离子内部状态的叠加转变为外部状态(即离子在“牢笼”内的振动状态)的叠加,甚至将一个离子的状态叠加转变为另一个离子的状态叠加。

与维因兰德的方法几乎恰好相反,阿罗什的囚禁物是被维因兰德当作“探针”的光子,而“探针”则类似于维因兰德的囚禁物,是一种被称为里德堡原子(Rydberg atom)的特殊原子,它的电子处于很高的能态上,从而使整个原子“发胖”到惊人的程度。比如阿罗什所用的铷(Rb)原子就“发胖”到了普通铷原子的500倍左右③。在阿罗什的方法中,囚禁光子所用的是以超导材料铌(Nb)制作的一对相距2.7厘米的球面镜,这对球面镜的工艺极为高超,构成了一个反射性质近乎完美的空腔(cavity)。光子在其中可以被反射十几亿次而不被吸收(在这过程中走过的总距离可以绕地球一圈)。在这些同样极不简

---

① 确切地说,最常用的离子阱有两种,一种叫做彭宁阱(Penning trap),另一种叫做保罗阱(Paul trap),他们的实现者分享了1989年的诺贝尔物理学奖。维因兰德所使用的是保罗阱。

② 边带冷却技术简单地说,是用能量为$\omega_i-\omega_v$(其中$\omega_i$为离子的内部能级差,$\omega_v$为离子在“牢笼”内的振动能级差)的光子,将处于振动能级$n>0$的离子激发到内部能级更高,但振动能级只有$n-1$的状态上(因为那样的光子只能将离子激发到那样的状态),然后让离子自行跃回原先的内部能级。由于离子在跃回过程中会优先维持振动能级不变,因此过程终了时离子的内部能级不变,振动能级却降为了$n-1$。重复这一过程(在必要时针对所需要的内部能级差调整光子能量),可以使振动能级最终降为基态$n=0$,从而达到冷却的目的。

③ 阿罗什所用的“发胖”后的铷原子的线度约为125纳米(nm),而普通铷原子的线度约为0.25纳米。

单的配置完成之后，阿罗什又通过特殊空腔中的电磁波，使作为“探针”的里德堡原子处于两个电子能态的叠加之中，并使之以可控制的速度穿越囚禁了光子的空腔。在这里，阿罗什做了另一个巧妙安排，使被囚禁光子的能量与里德堡原子所能吸收的能量稍稍错开，从而保证光子不会被里德堡原子所吸收（别忘了，这一整套方法的使命之一就是保障量子体系继续存在）。而更巧妙的是，尽管光子不会被吸收，它与里德堡原子的相互作用仍能对后者产生影响，改变后者那两个叠加能态间的相位。这样，阿罗什就可以通过研究穿越后的里德堡原子那两个叠加能态间的相位，而获得有关被囚禁光子的某些信息（比如光子的数目）。

上述两种方法的实现无疑都需要极高超的技术。不过，此类“工艺性”的工作要想获得诺贝尔奖，通常还需满足一个额外条件，那就是具有应用价值。此次获奖的工作很好地符合了这一条件，因为其所实现的“使得对单个量子体系的测量与操控成为可能的突破性实验方法”在理论与实用上都有着重要应用。

## 三、开放的应用

在理论上，对一个量子体系进行观测或操控，同时还让它继续存在，使得人们设计出了一些巧妙的实验，来观测量子体系状态演变的过程（以往的实验由于是“一锤子买卖”，对被观测体系具有“毁灭性”，从而无法做到这一点），甚至观测使一些物理学家深感困惑的量子体系的状态因为与外部环境的相互作用而往经典状态过渡的过程，其中包括对大名鼎鼎的“薛定谔的猫”（Schrödinger's cat）的生死过程的观测①。那样的实验已经有人做了。比如

① 当然，这是夸张的说法，事实上那猫被“掉包”成了一个量子体系，从而偏离了薛定谔拿猫“开涮”的本意——即通过引进作为宏观客体的猫，而彰显量子测量过程的佯谬性。不过包括诺贝尔委员会（Nobel Committee）提供的获奖工作介绍在内的大量资料和报道都已迫不及待地引入了“薛定谔的猫”一词。作为科普，我们姑且“从众”，但在这里略做说明，以图确切。

阿罗什本人的研究组就于2008年做了那样的实验，甚至将观测到的量子状态往经典状态过渡的过程制成了“影片”。

在实用上，此次获奖工作最引人注目的应用是在量子计算机领域。这是近年来被讨论得很多的领域，在乐观者看来，量子计算机若成为现实，对社会的变革将不亚于如今的计算机在过去几十年所带来的变革。不过，量子计算机的理论虽然美丽，面临的技术困难却极为巨大，其中一个很大的困难就是作为核心元件的量子体系必须能单个地、不受破坏地被测量与操控，而且各个量子体系的状态还必须能相互传递（就像经典计算机必须能在各元件间传递信息一样）。这个困难在过去几乎是难以克服的，此次的获奖工作却为之带来了曙光，比如维因兰德所实现的对状态叠加的操控，以及状态叠加在不同离子间的相互转变，就正是克服上述困难所需要的技术。这一点维因兰德本人也看得很清楚——事实上，他的研究组早已展开了这方面的探索，甚至在一定程度上构造出了量子计算机的雏形，实现了最简单的逻辑运算。一些其他实验组也正在积极努力之中。当然，这一切距离真正有实用价值的量子计算机还相差很远。

此次获奖工作的另一项很有价值的应用是建造超高精度的新型时钟。这一应用虽不像量子计算机那样富有未来色彩，所取得的进展却要扎实得多。维因兰德所供职的美国国家标准技术研究所正是这方面的“领头羊”。在这一应用中，用维因兰德所实现的方法囚禁起来的工作频率（即作为计时基础的两个能级之间的量子跃迁的频率）在光学波段的离子取代了传统原子钟所采用的工作频率在微波波段的铯（Cs）原子。目前，这种新型时钟已经达到了比传统铯原子钟高两个数量级的精度。在那样的精度下，哪怕从宇宙大爆炸之初开始计时，迄今的累计误差也只有区区几秒。

这些或已成为现实，或仍处于开放的想象空间里的应用，使此次的获奖工作有可能对未来科学与技术的发展产生深远影响。

## 附录：获奖者小档案



维因兰德　　阿罗什

- 维因兰德(David Wineland)：美国物理学家，1944 年 2 月 24 日出生于美国威斯康星州的密尔沃基(Milwaukee)，1970 年获哈佛大学(Harvard University)物理学博士学位，目前在美国科罗拉多州的国家标准技术研究所(National Institute of Standards and Technology)任职。维因兰德的主要研究方向为量子光学(quantum optics)及其应用。
- 阿罗什(Serge Haroche)：法国物理学家，1944 年 9 月 11 日出生于当时受法国控制的摩洛哥城市卡萨布兰卡(Casablanca)，1971 年获巴黎第六大学(Université Pierreet Marie Curie)的物理学博士学位，目前在法国巴黎(Paris)的法兰西公学院(Collègede France)任教。阿罗什的主要研究方向为量子光学及其应用。

2012 年 10 月 11 日写于纽约

# 第三部分　星际旅行漫谈

# 因为星星在那里

*Space, the final frontier!*

*StarTrek: The Next Generation*

试图挑战自然的人常会被问到为什么要用自己的生命去冒险。我有一位酷爱登山的朋友,一同在哥伦比亚大学(Columbia University)念研究生期间的某个夏天,他登上了北美洲的最高峰——海拔6 194米的麦金利峰(Mount McKinley)。我在系里遇见了刚从雪域高原回来的他。锐利的紫外线灼黑了他的皮肤,使我几乎认不出来,但一种敬意在我心中油然而生。我没有问他为什么要去登山,我知道登山家有一句震撼人心的名言:因为山在那里(Because it's there)。

小时候喜欢看星星,常可以看上几个小时不知倦怠。我知道天空中几乎每一颗小小的星星都要比我们脚下这个看似巨大的蓝色星球大上数百万倍,"大"与"小"竟以如此瑰丽的方式相互嵌套,那是何等的深邃和奇异啊!

30年前的1972年,人类向外太阳系发射了名为"先驱者10号"(Pioneer 10)的行星探测器。一年后又发射了它的姊妹探测器"先驱者11号"(Pioneer 11)。

它们已先后飞出了我们的太阳系(如果以冥王星轨道作为太阳系边界的话)。目前“先驱者 10 号”大约在距地球 120 亿千米之外，正向着 65 光年外的金牛座(Taurus)的毕宿五(Aldebaran)星飞去，以目前的速度计算将在约 200 万年后抵达。“先驱者 11 号”则将在约 400 万年后掠过天鹰座(Aquila)的一颗恒星。

200 万年对人类来说是一段过于漫长的时间：200 万年前人类还过着茹毛饮血的穴居生活；200 万年后当“先驱者 10 号”迎来自己孤独航程中第一缕耀眼的异星光芒时，人类也许早已在愚昧的战乱中成为了无言的化石。

登山家面对的是以人类微薄的体力去挑战大自然的伟岸，星际旅行家面对的则是以人类短暂的生命去跨越星际间几乎无限的距离。人类的平均寿命在过去几十年间虽然有所增长，但自然衰老依然是无可抗拒的规律。即使在基因图谱逐渐被揭开的今天，也没有迹象表明人类的寿命会在可预见的将来获得数量级上的延长。

从逻辑上讲，要让星际旅行家用短暂的生命去跨越近乎无限的时空，不外乎有两类方案：一类是从星际旅行家本身入手，设法在各种意义下延长其生命；另一类是从时空入手，设法利用或改变其结构，达到缩短空间距离或突破速度极限的目的。具体地讲，常见的设想有以下几种：

- 从星际旅行家本身入手的方案：
  - 用极低温“冷冻”的方法延长生命。
  - 用巨型空间站代替飞船，以群体繁衍的生命取代个体的生命。
  - 建造飞行速度接近光速的飞船，利用相对论的时间延缓效应达到延长生命的目的。
  - 将星际旅行家分解为基本粒子流或信息流以光速或接近光速的速度传播，并在目的地复现乘员。
- 从时空入手的方案：
  - 通过“虫洞”(wormhole)实现时空间的“捷径”(short-cut)旅行。
  - 通过“曲速引擎”(warp drive)实现“超光速”旅行。

“星际旅行漫谈”这个系列的文章将以目前所知的物理学规律为依据，来讨论其中的若干种方案，无论它们是出自科学家、工程师还是科幻小说家之手。

这些方案是人类探索璀璨星空的梦想的延续。

自远古以来这种梦想就以这样那样的方式存在着，历经无数的磨难和挫折，却从来不曾消失过。

因为人类的好奇心不可磨灭，因为星星在那里。

2002 年 7 月 24 日写于纽约

纪念“先驱者 10 号”发射 30 周年

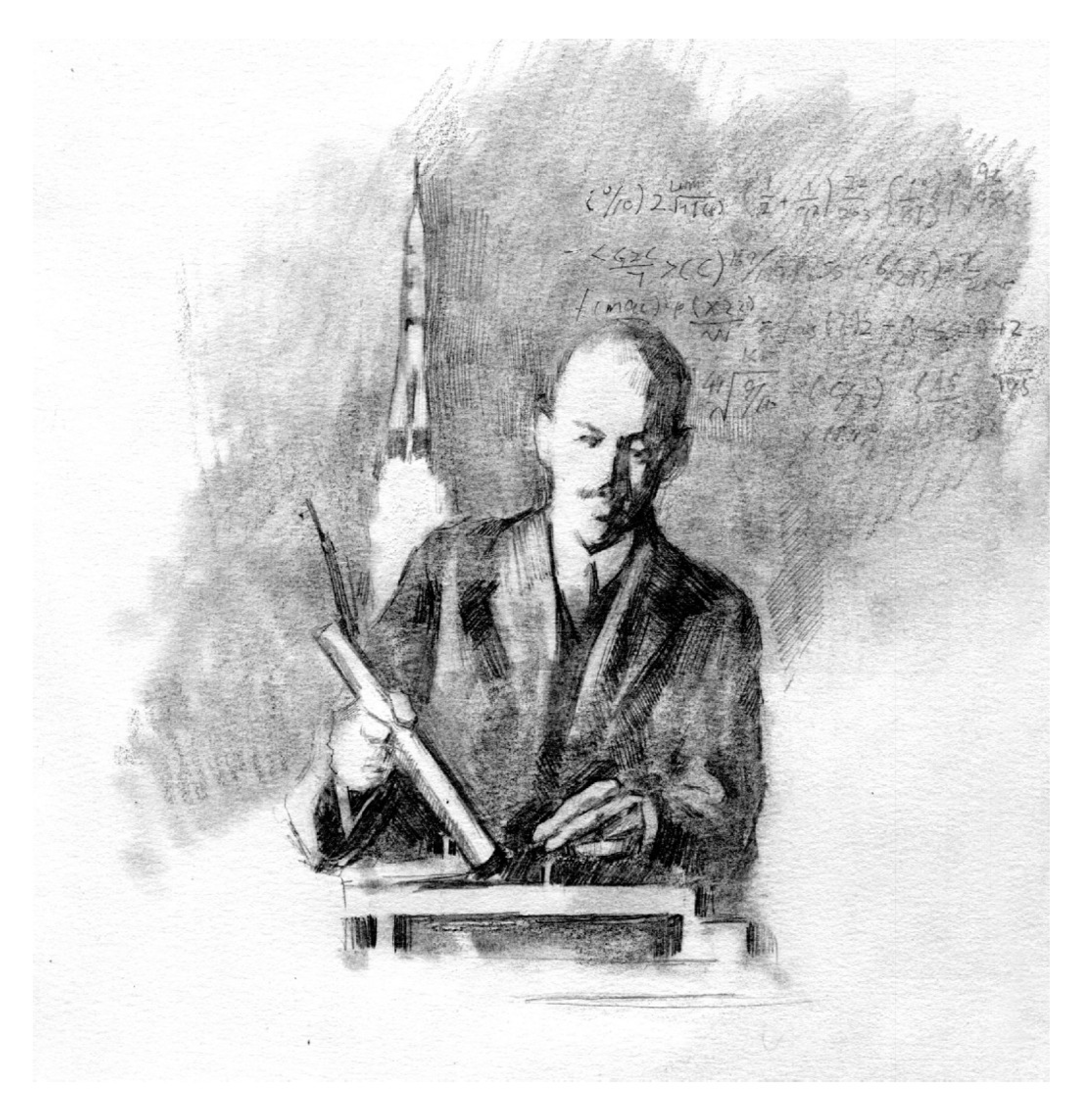
# 火箭：宇航时代的开拓者

## 一、引言

这个星际旅行系列原本是为了讨论未来的星际旅行技术而写的。不过今天却要来讨论一种比较“土”的技术：火箭。之所以讨论火箭，主要的原因有两个：一个是因为我国的第一艘载人飞船“神舟五号”即将发射①，在这个中国宇航员即将叩开星际旅行之门的时刻，我们这个系列不应缺席，也不应让火箭这位宇航时代劳苦功高的开拓者在这个系列中缺席。另一个是因为火箭虽然是一种不那么“未来”的技术，但在我和读者诸君能够看得到的未来，承载人类星际旅行之梦的技术很有可能仍然是火箭这匹识途的老马。

---

① 本文发表之后数小时，北京时间 2003 年 10 月 15 日早晨 9 时整，“神舟五号”飞船载着宇航员杨利伟从酒泉卫星发射中心发射升空。飞船升空 587 秒后与火箭分离，进入轨道倾角为 42.4 度、近地点高度为 200 千米、远地点高度为 350 千米的预定椭圆轨道。飞船飞行至第五圈时变轨进入高度为 343 千米的近地圆轨道。北京时间 2003 年 10 月 16 日早晨 6 时 23 分，飞船在环绕地球 14 圈后在内蒙古四子王旗北部的主着陆场安全着陆，不久杨利伟自主出舱。至此，我国第一次载人航天飞行取得圆满成功。杨利伟成为我国第一位进入太空的宇航员，我国成为继苏联与美国后第三个独立掌握载人航天技术的国家。“神舟五号”的发射是人类历史上的第 241 次载人航天飞行。杨利伟是人类历史上进入太空的第 952 人次。



绘画：张京

## 二、宇宙速度

火箭理论的先驱、俄国科学家齐奥尔科夫斯基（Konstantin Tsiolkovsky，1857—1935）有一句名言："地球是人类的摇篮。但人类不会永远躺在摇篮里，他们会不断探索新的天体和空间。人类首先将小心翼翼地穿过大气层，然后再去征服太阳周围的整个空间。"

星际旅行是一条漫长而坎坷的征途，人类迄今在这征途上所走过的部分几乎恰好就是"征服太阳周围的整个空间"，而这征途上的第一站也正是"穿过大气层"①。

在人类发射的航天器中，数量最多的就是那些刚刚"穿过大气层"的航天器——人造地球卫星，迄今已发射了数以千计。其中第一颗是1957年10月4日从苏联的拜克努尔航天发射场（Baikonur Cosmodrome）发射升空的"卫星一号"（Sputnik 1）。

从运动学上讲，这些人造地球卫星的飞行轨迹与我们随手抛掷的一块石头的飞行轨迹是属于同一类型的。我们抛掷石头时，抛掷得越快，石头飞得就越远，石头飞行轨迹的弯曲程度也就越小。倘若石头抛掷得如此之快，以致于飞行轨迹的弯曲程度与地球表面的弯曲程度相同，石头就永远也不会落到地面了②。这样的石头就变成了一颗环绕地球运转的小卫星，这一点早在牛顿（Isaac Newton，1642—1727）的《自然哲学的数学原理》（*Mathematical Principles of Natural Philosophy*）中就有过精彩的图示（图13）。一般地

---

① 大气层与行星际空间是连续衔接的，所谓"穿过大气层"指的是穿过厚度在百余千米以内的相对稠密的大气层。

② 当然，这里我们要忽略空气阻力，并且还要忽略地球表面的地形起伏。

讲，石头也好，卫星也罢，它们的飞行轨迹都是椭圆①。对于石头来说，如果抛掷得不够快，那它很快就会落到地面，从而我们就只能看到椭圆轨道的一个极小的部分，**那样的部分近似于一段抛物线**（感兴趣的读者请自行证明这一点）。

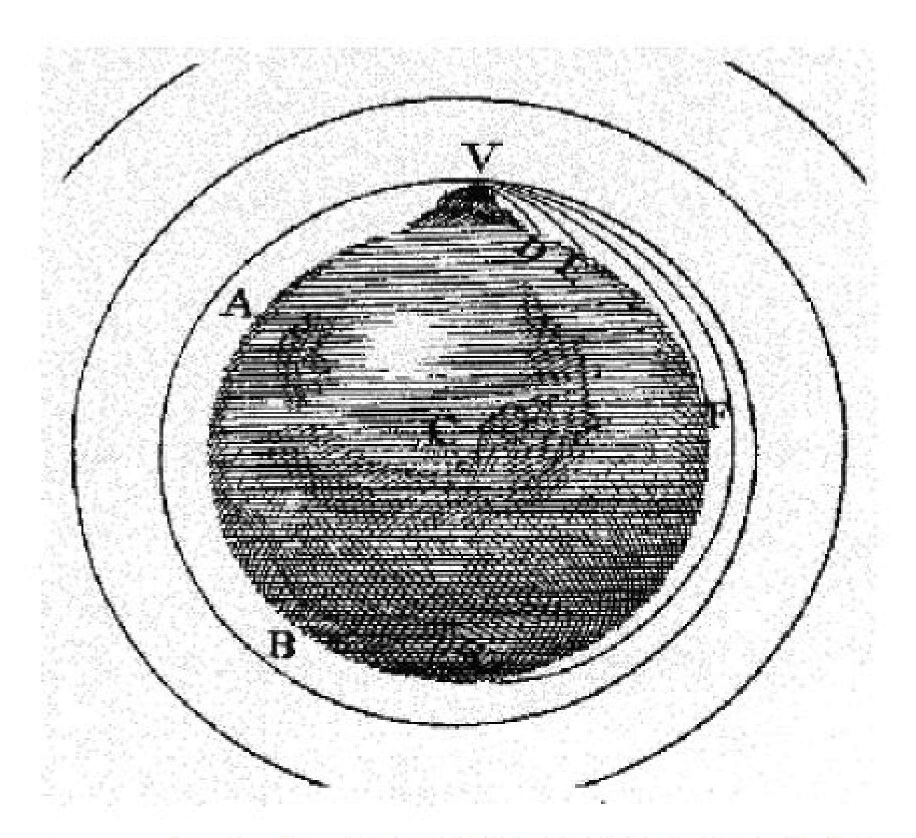



图 13　牛顿《自然哲学的数学原理》的插图

那么，一块石头要抛掷得多快才能不落回地面呢？或者说一枚火箭要能达到什么样的速度才能发射人造地球卫星呢？这个问题的答案很简单——尤其是对于圆轨道的情形。在圆轨道情形下，假如轨道的半径为 $r$，卫星的飞行速度为 $v$②，则维持卫星飞行所需的向心力为 $F=mv^2/r$（$m$ 为卫星质量），这一向心力来源于地球对卫星的引力，其大小为 $F=GMm/r$（$M$ 为地球质量）。由此可以得到 $v=(GM/r)^{1/2}$。假如卫星轨道很低（即轨道离地球表面很近），则 $r$ 约等于地球半径 $R$，由此可得 $v\approx7.9$ 千米/秒。这个速度被称为“第一宇宙速度”（first cosmic velocity），它是人类迈向星空所要达到的最低速度。

不过，细心的读者可能会从上面的计算结果中提出一个问题，那就是 $v=(GM/r)^{1/2}$ 随着轨道半径的增加反而在减小，这说明轨道越高的卫星飞行速度越小。但是直觉上，把东西扔得越高难道不应该越困难吗？再说，倘若把卫星

① 这里“卫星”指的是环绕地球运动的物体，其轨迹局限在有限区域内（否则的话，可能的轨迹将包括抛物线与双曲线）。同时我们还假定地球的引力场是一个严格的平方反比中心力场，且忽略任何其他星体的引力场。

② 确切地讲是指速度的大小，下文提到的“向心力”、“引力”等也往往指的是大小，请读者依据上下文自行判断。

发射得越高所需的速度反而越小，那么 $v\approx7.9$ 千米/秒这个“第一宇宙速度”岂不就不再是发射人造地球卫星所要达到的**最低速度**了？这些问题的出现，表明对于发射卫星来说，卫星的飞行速度并不是所需考虑的唯一因素。那么，还有什么因素需要考虑呢？答案是很多，其中最重要的一个是引力势能。事实上描述发射卫星困难程度的更有价值的物理量不是卫星的飞行速度，而是发射所需的能量，也就是把卫星从地面上的静止状态送到轨道上的运动状态所需提供的能量。因此我们改从这个角度来分析。在地面上，卫星的动能为零①，势能为 $-GMm/R$，总能量为 $-GMm/R$；在轨道上，卫星的动能为 $mv^2/2=GMm/2r$（这里运用了前面得到的 $v=(GM/r)^{1/2}$），势能为 $-GMm/r$，总能量为 $-GMm/2r$。因此发射卫星所需的能量为 $GMm/R-GMm/2r$。这一能量相当于把卫星加速到 $v=[GM(2/R-1/r)]^{1/2}$ 所需的能量。由于 $r>R$，这一速度显然大于 $v=(GM/r)^{1/2}\approx7.9$ 千米/秒（而且也符合轨道越高发射所需能量越多这一“直觉”）。这表明“第一宇宙速度”的确是发射人造地球卫星所需的最低速度，**只不过它表示的并不是卫星的飞行速度，而是火箭提供给卫星的能量所对应的等价速度**。在发射卫星的全过程中，火箭本身的飞行速度完全可以在任何时刻都低于这一速度。

上面的分析是针对圆轨道的，那么椭圆轨道的情况如何呢？在椭圆轨道上，卫星的飞行速度不是恒定的，分析起来要困难一些，但结果却同样很简单，卫星在椭圆轨道上的总能量仍然为 $-GMm/2r$，只不过这里 $r$ 表示所谓的“半长径”，即椭圆轨道长轴长度的一半。因此上面关于“第一宇宙速度”是发射人造地球卫星所需的最小（等价）速度的结论对于椭圆轨道也成立，是一个普遍的结论。

在人造地球卫星之后，下一步当然就是要把航天器发射到更远的地方——比方说月球上。为了实现这一步，火箭需要达到的速度又是多少呢？

---

① 这里参照系的原点取在地心，且忽略了由地球自转导致的卫星动能（因此而带来的误差小于1%）。

这个问题的答案也很简单，不过在回答之前先要对“更远的地方”做一个界定。所谓“更远的地方”，指的是离地心的距离远比地球半径（约为 $6.4\times10^3$ 千米）大，但又远比地球与太阳之间的距离（约为 $1.5\times10^8$ 千米）小。之所以要有后面这一限制，是因为在讨论中我们要忽略太阳的引力场①。由于航天器离地心的距离远比地球半径大，因此与发射前在地面上的引力势能相比，它在发射后的引力势能可以被忽略；另一方面，由于航天器不再做环绕地球的运动，其动能也就不再受到限制，最小可能的动能为零。（请读者想一想，这一动能是相对于什么参照系的？）因此发射后航天器的最小总能量近似为零。由于发射前航天器的总能量为 $-GMm/R$，因此需要由火箭提供给航天器的能量为 $GMm/R$，相当于把航天器加速到 $v=(2GM/R)^{1/2}\approx11.2$ 千米/秒的速度。这个速度被称为“第二宇宙速度”（second cosmic velocity），有时也被称为摆脱地球引力束缚所需的速度，它也是一个等价速度。

更进一步，倘若我们想把航天器发射得更远些，比方说发射到太阳系之外——就像本系列序言中所提到的“先驱者号”（Pioneer）探测器一样，火箭需要达到的速度又是多少呢？这个问题比前两个问题要复杂一些，因为所涉及的因素有地球与太阳两个星球的引力场，以及地球本身的运动。从太阳引力场的角度看，这个问题所问的其实就是在地球轨道所在处、相对于太阳的“第二宇宙速度”，即 $v=(2GM_S/R_{SE})^{1/2}$（其中 $M_S$ 为太阳质量，$R_{SE}$ 为地球轨道的半径，也即太阳与地球之间的距离）②。这一速度大约为 42.1 千米/秒。相对与第一、第二宇宙速度来说，这是一个很大的速度。但幸运的是，我们的地球本身就是一艘巨大的“宇宙飞船”，它环绕太阳飞行的速度约为 29.8 千米/秒。因此，如果航天器是沿着地球轨道运动的方向发射的，那么在远离地球时它相对于地球只要有 $v'=42.1-29.8=12.3$ 千米/秒的速度就行了。在地心参照

① 确切地讲是忽略太阳引力场中引力势能的变化。在这一限制之下其他行星的引力场也同样可以忽略。

② 这里我们忽略了地球轨道的微小椭率，而将之视为圆轨道。

系中，发射这样的一个航天器所需要的能量为 $mv'^2/2+GMm/R$（其中后一项为克服地球引力场所需要的能量，即前面计算过的把航天器加速到第二宇宙速度所需要的能量），相当于把航天器加速到 $v\approx16.7$ 千米/秒的速度。这一速度被称为"第三宇宙速度"（third cosmic velocity），有时也被称为摆脱太阳引力束缚所需的速度，它同样也是一个等价速度，而且还是针对在地球上沿地球轨道运动方向发射航天器这一特殊情形的。

以上三个"宇宙速度"就是迄今为止火箭技术所跨越的三个阶梯。在关于"第三宇宙速度"的讨论中我们看到，行星本身的轨道运动速度对于把航天器发射到遥远的行星际及恒星际空间是很有帮助的。这种帮助不仅在发射时可以大大减少发射所需的能量，而且对于飞行中的航天器来说，倘若巧妙地安排航线，也可以起到"借力飞行"的作用，比如"旅行者号"就曾利用木星的引力场及轨道运动速度来进行加速。

## 三、齐奥尔科夫斯基公式

在上节中我们讨论了为发射不同类型的航天器，火箭所要达到的速度。与火箭之前的各种技术相比，这种速度是很高的。在早期的科幻小说中，人们曾设想过用所谓的"超级大炮"来发射载人航天器。其中最著名的是法国科幻小说家凡尔纳（Jules Verne，1828—1905 年）的作品。凡尔纳在 1865 年发表的小说《从地球到月球》（*From the Earth to the Moon*）中曾经让三位宇航员挤在一枚与"神舟号"飞船的轨道舱差不多大的特制炮弹中，用一门炮管长达 900 英尺（约 300 米）的超级大炮发射到月球上去（最终没能击中月球，而成为了环绕月球运动的卫星）。不过，凡尔纳虽有非凡的想象力，却似乎缺乏必要的物理学及生理学知识。他所设想的超级大炮若真的在 300 米的炮管内把"炮弹"加速到 11.2 千米/秒（第二宇宙速度），则"炮弹"的平均加速度必须达到 200 000 米/秒$^2$ 以上，也就是 20 000$g$（$g\approx9.8$ 米/秒$^2$ 为地球表面的引力加速度）以上。但是脆弱的人类身体所能承受的最大加速度只有不到 10$g$。这

两者之间的巨大差异无疑是灾难性的，因此凡尔纳的炮弹虽然制作精致，乘坐起来却一点也不会舒适。不仅不会舒适，且有性命之虞。事实上，英勇的宇航员们在“炮弹”出膛时早就变成了肉饼，炮弹最后有没有击中月球对他们都已不再重要了。而且若炮弹真的击中月球的话，其着陆方式属于所谓的“硬着陆”，就像陨石撞击地球一样，着陆时的速度差不多就是月球上的第二宇宙速度（约为 2.4 千米/秒），相当于在地球上从比珠穆朗玛峰还高 30 倍的山峰上摔到地面，这无疑是要把肉饼进一步摔成肉酱。

因此对于发射航天器（尤其是载人航天器）来说，很重要的一点就是航天器的加速过程必须发生在一个较长的时间里（减速过程也一样）。但是加速过程持续的时间越长，在加速过程中航天器所飞行的距离也就越大。以凡尔纳的超级大炮为例，倘若炮弹的加速度小于 $10g$，则加速过程必须持续 100 秒以上，在这段时间内炮弹飞行的距离在 500 千米以上。炮弹的加速度越小，这段距离就越大。由于炮弹本身没有动力，因此这段距离必须都在炮管内。这就是说，凡尔纳超级大炮的炮管起码要有 500 千米长！建造这样规模的大炮显然是很困难的，别说凡尔纳时代的技术无法办到，即使在今天也是申请不到经费的。因此航天器的发射必须另辟蹊径①。火箭便是一种与凡尔纳大炮完全不同但却非常有效的技术手段。

火箭是一种利用反冲作用推进的飞行器，即通过向与飞行相反的方向喷射物质而前进的飞行器。从物理学上讲这种飞行器所利用的是动量守恒定律。下面我们就来对火箭的飞行动力学作一个简单分析。

假设火箭在单位时间内喷射的物质质量为 $-\mathrm{d}m/\mathrm{d}t$（$m$ 为火箭质量，$\mathrm{d}m/\mathrm{d}t<0$），喷射物相对于火箭的速度大小为 $u$（方向与火箭飞行方向相反），则在时间间隔 $\mathrm{d}t$ 内，火箭的速度会因为喷射而得到一个增量 $\mathrm{d}v$。依据动量守恒定律，在火箭参照系中可以得到

① 类似于凡尔纳大炮那样的装置在表面引力较弱的星球——比如月球——上建造起来就会容易许多，因此曾有人设想它可以成为未来月球基地的航天器发射装置。

$$m\mathrm{d}v = -u\mathrm{d}m$$

对上式积分并注意到火箭的初速度为零，便可得到

$$v = u\ln(m_i/m_f)$$

其中 $m_i$ 与 $m_f$ 分别为火箭的初始质量及推进过程完成后的质量（显然 $m_i > m_f$）。这一公式被称为齐奥尔科夫斯基公式（Tsiolkovsky formula），它是由上文提到过的俄国科学家齐奥尔科夫斯基发现的，时间是 1897 年，那时候的天空还是人类的“禁地”，连飞机都还没有上天①。齐奥尔科夫斯基因为在航天领域的一系列卓越的开创性工作，而被许多人尊称为“航天之父”（father of astronautics）或“火箭之父”（father of rocketry）。

从齐奥尔科夫斯基公式中我们可以看到，火箭所能达到的速度可以远远地高于喷射物的喷射速度。这一点是很重要的，因为这意味着我们可以通过一种较低的喷射速度来达到航天器所需要的高速度，这在技术上远比直接达到高速度容易得多。从某种意义上讲，凡尔纳的超级大炮之所以没能成为一种载人航天器的发射装置，正是因为它试图直接达到航天器所需要的高速度。

但是火箭虽然能够达到远比喷射物喷射速度更高的速度，为此而付出的代价却也不小，因为火箭所要达到的速度越高，它的有效载荷就必须越小。这一点从齐奥尔科夫斯基公式中可以很容易地看到。我们可以把公式改写为 $m_f = m_i\exp(-v/u)$，由此可见，火箭的飞行速度 $v$ 越高，它的有效载荷（$m_f$ 中的一部分）也就越小。假如我们想用 $u = 1$ 千米/秒的喷射速度来达到第一宇宙速度（即将有效载荷送入近地轨道），则 $m_f/m_i \approx 0.000\,37$，也就是说一枚发射质量为 1 000 吨的火箭只能让几百千克的有效载荷达到第一宇宙速度，这样的效率显然是太低下了。

为了克服这一困难，齐奥尔科夫斯基提出了多级火箭的设想。多级火箭

① 这一公式的正式发表是在 1903 年，与莱特兄弟（Wright brothers）的飞机同一年。另外，新近发现的一些史料表明，英国皇家军事学院（Royal Military Academy）的科学家早在 1813 年就得到过类似的结果。

的好处是在每一级火箭的燃料用尽后可以把该级火箭的外壳抛弃掉，从而减轻下一级火箭所负载的质量。在理论上，火箭的级数越多，运载效率就越高，不过在实际上，超过三级的火箭其技术复杂性的增加超过了运载效率上的优势，使用起来得不偿失。因此，目前我们使用的火箭大都是三级火箭。即便使用多级火箭，航天飞行的消耗依然是惊人的，通常一枚发射质量为几百吨的火箭只能将几吨的有效载荷送入近地轨道，比如发射“神舟号”飞船的长征二号F型火箭发射质量约为480吨，近地轨道的有效载荷约为8吨。

## 四、接近光速

前面说过，这个星际旅行系列主要是为了讨论未来的星际旅行技术而写的，因此，在这里我们也要把目光放远些，看看上节讨论的火箭动力学在火箭速度持续提高，乃至接近光速时会如何。截至2013年7月，人类发射的航天器中飞得最远的是1977年9月5日发射的“旅行者一号”（Voyager 1）。经过近36年的漫长飞行，它已经飞到了离太阳约187亿千米处，远远超出了太阳系已知最外围的行星——海王星，或曾经最外围的行星——冥王星——的轨道。但是，这个距离跟离太阳最近的恒星——半人马座比邻星（Proxima Centauri）——的距离相比，还不到万分之五。由此可见，人类要想走得更远，必须要有更快的航天器。在齐奥尔科夫斯基公式中火箭的速度是没有上限的，通过提高喷射物的喷射速度，通过增加火箭质量中喷射物所占的比例，火箭在原则上可以达到任意高的速度。但是，这一点显然是错误的，因为物体的运动速度不可能超过光速，这是相对论的要求①。这表明，当火箭的运动速度接近光速时，齐奥尔科夫斯基公式将不再成立。那么，有没有一个比齐奥尔科

① 在理论与实验上都有迹象表明，在特定的条件及特定的含义下，运动速度超过光速并非绝对不可能，但这种超光速并不像许多科普爱好者所认为的那样，是推翻了相对论。

夫斯基公式更普遍的公式，在火箭运动速度接近光速时仍成立呢？这就是本节所要讨论的问题。

首先，简单的答案是：这样的公式是存在的。事实上，这样的公式不仅存在，而且并不复杂，因此我们干脆在这里把它推导出来，以满足大家的好奇心。这一推导所依据的基本原理仍然是动量守恒定律，我们也仍然在火箭参照系中计算火箭速度的增量。这里要补充说明的是，所谓火箭参照系，指的是所考虑的瞬间与火箭具有同样运动速度的惯性参照系(因此在不同的时刻，火箭参照系是不同的)。我们用带撇的符号表示火箭参照系中的物理量(这是讨论相对论问题的惯例)。与上节的讨论相仿，假设火箭在单位时间内喷射的物质质量为$-\mathrm{d}m'/\mathrm{d}t'$($m'$为火箭质量，$\mathrm{d}m'/\mathrm{d}t'<0$)，喷射物相对于火箭的速度大小为$u$(方向与火箭飞行方向相反)，则在时间间隔$\mathrm{d}t'$内，火箭的速度会因为喷射而得到一个增量$\mathrm{d}v'$。依据动量守恒定律，在火箭参照系中可以得到

$$m'\mathrm{d}v' = -u\mathrm{d}m'$$

这里$\mathrm{d}m'$为喷射物的相对论质量(运动质量)，这一公式对于$u$接近甚至等于光速的情形也成立①。在非相对论的情形下，上面所有带撇的物理量都等于静止参照系(地心参照系)中的物理量，因此对上述公式可以直接积分，这种积分的含义是对上式中的速度增量进行累加。但在相对论中，速度合成的规律是非线性的，把这些在不同时刻——因而在不同参照系中——的速度增量直接累加是没有意义的，因此上述速度增量必须先换算到静止参照系中才能积分。

运用相对论的速度合成公式，$\mathrm{d}v'$所对应的静止系中的速度增量为

$$\mathrm{d}v = \frac{\mathrm{d}v' + v}{1 + \frac{v\mathrm{d}v'}{c^2}} - v = \left(1 - \frac{v^2}{c^2}\right)\mathrm{d}v'$$

将这一结果与在火箭参照系中所得的关于$\mathrm{d}v'$的公式联立可得

① 假如$u$等于光速，则$\mathrm{d}m'$理解为$\mathrm{d}E'/c^2$($E'$为喷射物的能量)。

$$\frac{\mathrm{d}v}{1-\frac{v^2}{c^2}}=-u\frac{\mathrm{d}m'}{m'}$$

对这一公式积分，并进行简单处理，便可得到

$$v=c\tanh\left(\frac{u}{c}\ln\frac{m_{\mathrm{i}}}{m_{\mathrm{f}}}\right)$$

其中火箭的初始质量 $m_{\mathrm{i}}$ 与推进过程完成后的质量 $m_{\mathrm{f}}$ 都是在火箭参照系中测量的。这就是齐奥尔科夫斯基公式在相对论条件下的推广。对于低速运动的火箭，$(u/c)\ln(m_{\mathrm{i}}/m_{\mathrm{f}})\ll 1$，因而 $\tanh[(u/c)\ln(m_{\mathrm{i}}/m_{\mathrm{f}})]\approx(u/c)\ln(m_{\mathrm{i}}/m_{\mathrm{f}})$，上述公式退化为普通的齐奥尔科夫斯基公式。由于对于任意 $x$，$\tanh x<1$，因此由上述公式给出的速度在任何情况下都不会超过光速，从而符合相对论的要求。

上述公式的一个特例是 $u=c$ 的情形，即喷射物为光子(或其他无质量粒子)的情形。这种火箭常常出现在科幻小说中，通常是以物质与反物质的湮灭作为动力来源。对于这种情形，上述公式简化为：$v=c(m_{\mathrm{i}}^2-m_{\mathrm{f}}^2)/(m_{\mathrm{i}}^2+m_{\mathrm{f}}^2)$。如果将火箭 90％的质量转化为能量作为动力，火箭的飞行速度可以达到光速的 99％。

## 五、飞向深空

宇宙的浩瀚是星际旅行家们所面临的最基本的事实。即使能够达到接近光速的速度，飞越恒星际空间所需的时间仍然是极其漫长的。比如从太阳系出发，到银河系中心大约要 3 万年，到仙女座星云(Andromeda Galaxy，也称为 M31，为河外星系)大约要 220 万年，到室女座星系团(Virgo，为河外星系团)大约要 6 000 万年……相对于人类弹指一瞬的短暂生命来说这些时间显然是太漫长了。但是且慢悲观，因为我们还有一个因素可以依赖，那就是相对论的时钟延缓效应。在相对论中运动参照系中的时间是由所谓的“本征时间”来表示的，它与静止参照系中的时间之间的关系为

$$\tau=\int\left(1-\frac{v^2}{c^2}\right)^{1/2}\mathrm{d}t$$

把这个公式运用到火箭参照系中，$\tau$ 就是宇航员所感受到的时间流逝。很显然，火箭的速度越接近光速，宇航员所感受到的时间流逝也就越缓慢。考虑到这个因素，宇航员是不是有可能在自己的有生之年到银河系中心、仙女座星云、甚至室女座星系团去旅行呢？下面我们就来计算一下。

我们考虑一个非常简单的情形，即火箭始终处于匀加速过程中。当然这个匀加速度是在火箭参照系中测量的。为了让宇航员有“宾至如归”的感觉，我们把加速度选为与地球表面的重力加速度一样，即 $g$。用数学语言表示：

$$\frac{\mathrm{d}^2x'}{\mathrm{d}t'^2}=g$$

把这一加速度变换到静止参照系（地心参照系）中可得

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2}=\left(1-\frac{v^2}{c^2}\right)^{3/2}g$$

由此积分可得

$$x=\frac{c^2}{g}\left[\left(1+\frac{g^2t^2}{c^2}\right)^{1/2}-1\right]$$

只要加速的时间足够长（即 $gt\gg c$），上式可近似为 $x\approx ct$。这表明在地心参照系中，经过长时间加速后飞船基本上是以光速飞行的。但是我们感兴趣的是宇航员所经历的时间，即“本征时间”$\tau$，这是很容易利用上式——$\tau$ 的定义——计算出的，结果为（请读者自行验证）

$$\tau=\frac{c}{g}\operatorname{arcsinh}\left(\frac{gt}{c}\right)$$

我们可以从 $\tau$ 和 $x$ 的表达式中消去 $t$，由此得到

$$\tau=\frac{c}{g}\operatorname{arcsinh}\left\{\left[\left(1+\frac{gx}{c^2}\right)^2-1\right]^{1/2}\right\}$$

如果 $x\ll c^2/g$（约 1 光年），即飞行距离远小于 1 光年，上式可近似为：$\tau\approx(2x/g)^{1/2}$，这正是我们熟悉的非相对论匀加速运动的公式。如果 $x\gg c^2/g$，即飞行距离远大于 1 光年，上式可以近似为 $\tau\approx(c/g)\ln(2gx/c^2)$。下面我们将

只考虑这种情形。考虑到抵达一个目的地后，通常还要做一些考察研究、拍照留念的事情，因此火箭不能一味加速，而必须在航程的后半段进行减速，从而旅行所需的时间应当修正为（最右侧表达式中 $\tau$ 以年为单位，$x$ 以光年为单位）

$$\tau = \frac{2c}{g}\ln\left(\frac{gx}{c^2}\right) \sim 2\ln x$$

由这一公式不难看到：倘若旅行的目的地是银河系的中心，$x=30\ 000$ 光年，则 $\tau \sim 20$ 年。这就是说，在宇航员看来，仅仅 20 年的时间，他就可以到达银河系的中心，即使考虑到返航的时间，前后也只需 40 年的时间，他就可以衣锦还乡了。这就是相对论的奇妙结论！只不过，当他回到地球时，地球上的日历已经翻过了整整 6 万年，他的孙子的孙子的孙子……（如果有的话）都早已长眠于地下了①。

运用同一公式，我们还可以计算出到达仙女座星云所需的时间约为 29 年，到达室女座星系团所需的时间约为 36 年……（在这里，读者们对于对数函数的增长之缓慢大概会有一个深刻印象吧。）倘若一个宇航员 20 岁时坐上火箭出发，如果他可以活到 80 岁，那么在他有生之年（不考虑返航——“壮士一去兮不复返”），他可以到达 10 000 000 000 000（10 万亿）光年远的地方。这个距离已经远远远远地超过了可观测宇宙的线度。因此，这样一位宇航员在其有生之年可以到达宇宙中任意远的地方！

由此看来，星际旅行似乎并不像人们渲染的那样困难。倘如此，则我们也就不必费心讨论什么虫洞（wormhole）和生命传输机（transporter）了，直接坐上火箭遨游太空就是了。事情当然并不如此简单，别忘了在我们的计算中火箭是一直在加速的（否则的话，那个帮了我们大忙的对数函数就会消失），那样的火箭所耗费的能量是惊人的（究竟要耗费多少能量呢？运用本文给出的结

① 这类结果早年曾引起过争议，并被称为“时钟佯谬”（clock paradox），但其实并无佯谬可言，感兴趣的读者可参阅拙作“关于时钟佯谬”（已收录于本书）。

果，读者可以自己试着计算一下)[①]。不过这种能量耗费所带来的困难比起建造虫洞所面临的困难来终究还是要小得多。因此，运用那样的火箭探索深空也许真的会成为未来星际旅行家们的选择。唯一的遗憾是，他们只要走得稍远一点，我们就没法分享他们的旅行见闻了。

因为相对论只保佑他们，不保佑我们。

2003 年 10 月 14 日写于纽约

2013 年 7 月 13 日最新修订

① 需要提醒读者的是，这种速度极其接近光速的火箭将会遇到的一个我们未曾提及的问题，那就是：它所经过的星际空间中的所有物质——哪怕细微到基本粒子——相对于火箭都具有极高的能量，从而有可能造成极大的危害。

# 生命传输机[①]

看过科幻电视连续剧《星际迷航》(*Star Trek*)的人可能对剧中的生命传输机(Transporter)留有深刻的印象。需要进入别的飞船或在星球上着陆的飞船乘员站在生命传输机的控制室中,随着操作人员的一句“Energize”的口令,乘员的身体渐渐分解成了一片闪烁的粒子,从控制室中悄然消失;几乎与此同时,在传输目的地,一个粒子团魔术般地出现,并渐渐变得明亮起来,最终完整地复现出了飞船乘员(图 14)。整个分解和复合的过程只需几秒钟。据说《星际迷航》的编导们最初设计这么一个生命传输机是为了省钱,因为当时摄制组的经费负担不起拍摄星际飞船在星球表面着陆所需的特技过程。

像生命传输机那样的概念使许多人都感到了兴趣。念中学时我曾翻过一本由美国学者霍夫施塔特(Douglas R. Hofstadter)和丹尼特(Daniel C. Dennett)撰写的名为《心我论》(*The Mind's I*)的书,一开头就提到了类似于生命传输机

① 本文曾发表于《科学画报》2003 年第 10 期(上海科学技术出版社出版)。



图 14　生命传输机

的装置，由此展开了许多生命哲学方面的讨论。对研究星际旅行的人来说，像生命传输机那样的装置是让脆弱而短暂的生命以基本粒子的形式跨越星际间严酷的环境和近乎无限的时空尺度的理想手段。

《星际迷航》播映之后还出版了一本《技术手册》（*Technical Manual*），替剧中用到的许多新技术和新概念作了书面描述。从《技术手册》上看，《星际迷航》中的生命传输机是直接将组成原生命体的基本粒子传输到目的地进行复现的。按照我们对微观世界的了解，这是不必要的。因为依据量子力学的基本原理，同一类型的基本粒子彼此间是完全相同的。因此在使用生命传输机的过程中，组成生命体的那些基本粒子本身是否直接被传输到目的地其实并不重要，因为那些基本粒子本身并没有任何特殊性。真正需要传输的只是有关生命微观组成的完整信息[①]。只要有了这些信息，通过什么途径，从什么地方获取复现生命体所需的基本粒子是无关紧要的。事实上，生命虽然奥妙，但组成生命体的那些基本粒子——注意不是分子，而是基本粒子——本身据我们所知在宇宙间是普遍存在的。因此，如果有一天星际旅行家们真的建造出了像生命传输机那样的装置，我们所要做的将只是设法把接收和复现装置送

① 在后文中将会提到，对这里所说的“完整”两字不宜理解得过于绝对。

到目的地（《星际迷航》中连这些装置也省略了，看来经费的确是比较紧张），此后两地之间的旅行在原则上就可以像今天人们所熟悉的电波通信那样快捷和“方便”了。

那么像生命传输机那样能够把生命分解为基本粒子，并在异地完整复现的装置在物理上是否可以实现呢？如果可以实现，它的作用过程是否会像人们在《星际迷航》中所看到的那样呢？这些就是本文所要讨论的问题。至于生命传输机所引发的有关生命哲学方面的思考则不在本文的考虑之列，感兴趣的朋友可以去看看《心我论》或其他类似的书。

按照前面的介绍，生命传输机在物理上能否实现的一个关键的环节，就在于能否获得有关生命微观结构的完整信息。我们不妨回想一下，在宏观世界里如果我们要复制一样东西，比方说一件家具，该怎么做？通常我们会从各个角度对所要复制的家具进行观察，研究它的材料，分析它各部件的拼合方式，如此等等。从物理学的角度讲，所有这些都是对被复制的物体进行观测，复制过程所需的信息就来源于这些观测。这些观测所需达到的细微程度则显然与复制本身所需达到的精密程度密切相关。对于家具而言，人们关心的是它的外观、手感、强度等性质，复制物只要在这些性质上做到与原件难以区分就可以了。由于这些性质都是宏观性质，有关它们的信息都是宏观信息，因此为复制家具所需的观测是宏观意义上的观测，这样的观测在物理学上是没有任何原则性困难的。

那么复制生命的情况又如何呢？这里所说的复制生命不是今天大家正在热议的克隆（clone），克隆所复制的只是生命的躯壳，而我们讨论的是真正地、全息意义上的生命复制。这种复制不仅包括躯壳，还必须包括记忆、意识、情感、智慧等原生命体所具有的全部重要特征。这里我们遇到的第一个巨大的困难就是我们并不清楚生命——尤其是像人类这样的“高等”生命——的全部奥秘，比方说我们迄今还不了解意识的物理起源。我们不清楚人的意识以及其他许多深层功能的存在究竟是依赖于人体在哪个物质层次上的结构，是原子、分子层次？还是细胞层次？亦或干脆就是一种独立的存在？依据答案的

不同，为传输生命所需获得的有关生命结构的信息，以及在传输和复现生命过程中所需使用的物质基元（building block）将会有所不同。

很明显，在没有找到这些问题的真正答案之前是无法对复制生命的可行性做出准确判断的。不过从星际旅行的角度讲，如果生命传输机所需传输的是细胞（或细胞以上的组织），那么由于细胞本身就是一种初等的生命，在星际间的环境和时间跨度上维持它们与直接让人进行星际旅行所面临的困难也许只有程度上的差别，从而生命传输机对于星际旅行的价值就要大打折扣。本文将不讨论这种类型的生命传输机（《星际迷航》中的生命传输机显然也不是这一类型的）。另一方面，如果复制生命需要涉及非物质的东西（比方说如果意识是物质以外的独立存在），那么我们目前显然尚不具备讨论这一问题的物理学依据。

因此本文所要——或者说所能够——讨论的只有一种情形：即对生命的复制是在原子、分子或其他基本粒子层次上进行的。这也是生命传输机对星际旅行来说具有最大价值的情形（《星际迷航》中的生命传输机就属于这一类型）。因为正如前面所说，同一类型的基本粒子（或简单的粒子组合如原子、分子）在量子力学意义上是全同的，而且在这一层次上物质的组元（质子、电子等）在宇宙中是普遍存在的，这就使得直接传输组成生命的物质（以及维持这种物质）成为不必要，从而大大简化了生命传输机的结构。对于这种类型的生命传输机，只要我们能获得有关生命微观结构的完整信息，它的制造以及它在星际旅行中的使用至少在理论上就具有了相当大的可能性。

因此问题归结为我们是否有可能获得有关生命微观结构的完整信息。

在讨论如何获取有关生命微观结构的完整信息之前，让我们先来估计一下这种信息的数量，以便大家有个概念。人体大约由一万亿亿亿（$10^{28}$）个原子组成。假如对这一结构中每个原子的描述（包括它与周围原子的连接方式）平均需要 100 比特（byte）的信息，那么有关生命微观结构的完整信息大约有 $10^{21}$GB（一个 GB 约等于 10 亿比特）。$10^{21}$GB 的信息是个什么概念呢？打个比方吧，这样数量的信息，如果用容量为 100GB 的计算机硬盘来储存，大约需

要 1 000 亿亿张硬盘。这些硬盘如果摆放起来的话,足以覆盖整个地球表面(不分陆地海洋)100 遍!

传输和储存如此大量的数据本身无疑也是一个很大的挑战,但这种挑战相对于复制生命所面临的全部复杂性来说只不过是冰山之一角!

复制生命的真正复杂性来自这样一个事实:那就是获取一个体系微观上的完整信息在物理学上远不是一件轻而易举的事情,它和复制家具所涉及的获取体系的宏观信息有着本质的差别。这一差别来自于今年已逾百岁"高龄"的量子力学。一百多年前,自伽利略(Galileo Galilei)和牛顿(Isaac Newton)以来岿然屹立已达数百年之久的经典物理学大厦如同一串精巧的多米诺骨牌,被一朵"物理学晴朗天空中的小小乌云"——黑体辐射问题——撞了一下腰,竟尔轰然倒塌。所幸的是物理学本身就像浴火重生的火凤凰,从灰烬中脱胎出了一个崭新的领域,那便是量子力学。但是,对钟情于生命传输机的星际旅行家们来说,不幸的是:获取一个体系微观上的完整信息的美好愿望却被无情地压在了经典物理学的那片厚厚的废墟下面……

量子力学的出现导致了物理理论及其描述自然的总体方式的彻底变革。在量子力学中,对一个物理体系的描述由所谓的"波函数"(wave function)来表示①。许多传统的经典物理学概念——比如粒子所在的位置、粒子的运动速度,等等——失去了经典物理学赋予它们的实在性。量子力学诞生之后,尤其是著名的"不确定性原理"(uncertainty principle)提出前后,物理学家们对这一理论的内涵、它的自洽性和完备性等问题进行了长时间激烈的争论。那些争论大大澄清和加深了人们对许多量子力学基本概念的理解。从那些让物理学获益良多的争论中衍生出了许多全新的分支领域,其中的一个叫做量子力学测量理论,它是我们讨论获取一个体系微观上的完整信息的理论依据。

---

① 确切地说,在量子力学中,对一个物理体系的描述体现在所谓的"状态"(state)上,"波函数"是状态在具体表象——比如坐标表象——下的函数表示。

自测不准原理提出以来，物理学家们对量子力学测量理论的研究已经进行了整整四分之三个世纪。如果注意到这种研究是在量子力学的基本数学框架未出现重大变动的情况下进行的，并且有20世纪几乎所有最伟大的物理学家——比如爱因斯坦（Albert Einstein）、玻尔（Niels Bohr）、海森伯（Werner Heisenberg）、玻恩（Max Born）、薛定谔（Erwin Schrödinger）等——的积极参与，却直到今天也没能形成一个被普遍认可的理论，这在科学史上是颇为罕见的。量子力学在概念层次上的微妙性由此可见。量子力学的初学者们常常被告诫："如果初学量子力学就觉得明白了，那你一定是没有理解它。"在量子力学炽热发展的时期，新的理论模型层出不穷。据说当时评判一个新理论是否正确的"标准"之一就是看这个理论是否足够"疯狂"，如果不是，那它一定是错的！全面地讨论量子力学测量理论远远超出了本文的范围。不过，值得庆幸的是虽然并不存在一个被普遍认可的测量理论，但分歧主要是集中在对理论的诠释上，物理学家们对测量理论的一些主要结论还是相当程度的共识的。简单地说，量子力学测量理论有别于经典测量理论的一个最基本的特点就是：观测过程本身对被观测体系造成的干扰是不可忽略的。用一句许多量子物理学家喜爱的俗语来表述就是：在量子力学这部大戏中，观测者既是观众也是演员。

量子测量理论的这一特点对获取有关生命微观结构的完整信息会造成一个很棘手的问题，那就是体系的微观状态经过一次测量就会发生变化。而状态一变，此后的测量所获得的就不再是关于体系原先微观状态的信息了。这就是说对一个体系的微观状态只能进行一次有效的测量。当然，"一次测量"在逻辑上并不意味着就只能得到"一点点"信息，我们也许可以期盼某种非常"聪明"的测量方法，一次就可以得到一个量子体系的全部信息。不幸的是，量子力学测量理论的另一个著名的结论就是：有一些可观测量是相互排斥，从而不可能在一次测量中同时获得精确结果的。换句话说，对一个量子体系的单次测量所能得到的信息往往注定只能是不完整的！

在研究普通的量子体系——比如氢原子——时这一点并不造成实质的困

难,因为自然界中所有的氢原子都是一样的。我们可以对许多氢原子进行独立的测量,然后对结果进行综合分析。这正是对一个量子体系进行测量的标准方法。事实上在考虑量子力学测量问题时人们通常引进所谓的“系综”(ensemble)——即大量全同体系的集合——的概念,对一个量子力学体系的测量事实上是针对系综中各个全同体系进行大量的独立测量。这些独立测量的结果的统计分布由体系的波函数所描述①。反过来,通过选择适当的待测物理量或物理量的组合,对一个系综中各个全同体系进行充分多的独立测量,从测量结果中原则上也可以反推出体系的波函数来。而波函数一旦确定,在量子力学意义上也就获得了有关体系微观结构的完整信息。

很明显,把这套理论用到我们所讨论的获得有关生命微观结构的完整信息的问题上来就会陷入一种“先有鸡还是先有蛋”的循环之中。因为按照上述理论,为了获取关于某个生命体微观结构的完整信息,必须先制备一个关于这一生命体的系综。但是生命体不像氢原子那样具有微观全同性,自然界中根本就不存在关于生命体的系综。这就意味着要想制备一个关于生命体的系综,我们必须自行复制生命体。而为了能够复制一个生命体,我们就需要先知道关于该生命体微观结构的完整信息。

绕了一圈我们依然两手空空。

因此,获得有关生命微观结构的完整信息按照我们今天对量子力学规律的理解是不可能的。如果复制生命——从而制造生命传输机——果真严格依赖于有关生命微观结构的完整信息,那它就同样是不可能的。不过“幸运”的是,虽然我们并不清楚生命——包括记忆、意识、情感、智慧等全部内涵——对微观结构的确切依赖程度,但这种依赖必定带有某些程度的模糊性。也就是说微观状态的某些程度的改变不会影响生命的任何本质特征。比方说头上缺几根头发,皮肤上多一两点色斑,身上少几个细胞等所对应的微观状态的差

① 这里所说的系综理论只是量子力学测量理论所涉及的若干种诠释中的一种,但可以算是最直接对应于量子力学数学体系的诠释。

异显然都不会妨碍所复制生命的有效性。因此我们所需回答的问题可以弱化为：考虑到所有可被允许的模糊性，是否有可能获得复制生命所必须的微观信息？遗憾的是，对这一问题我们目前只能用一个双重的“无可奉告”来回答。因为我们既不清楚“可被允许的模糊性”的确切含义，也没有对量子力学测量理论研究到足以回答这类问题的透彻程度。我们比较有把握的结论是：在简单意义上精确复制生命——即复制生命的全部微观结构——的生命传输机是不可能制造的。

最后我们再讨论一下如果生命传输机存在，它的工作情形是否会像图 14 所示的那样干净利落，在几秒钟之内点尘不惊地完成复制过程。当然，我们不可能讨论生命传输机的具体工作方式，我们只想来计算一下把一个人分解为基本粒子或由基本粒子复合成一个人所需吸收或释放的能量。假如生命传输机对人体的分解和复合是在亚原子——即质子、中子、电子等——的层次上进行的，那么人体将会被分解为大约 10 万亿亿亿（$10^{29}$）个亚原子粒子（比上文提到的原子数目多一个数量级左右）。由于平均每个亚原子粒子的结合能约为 1 兆电子伏特（1Mev），因此分解（复合）过程所需吸收（释放）的能量大约为 1 亿亿焦耳（$10^{16}$J），这相当于 100 万吨 TNT 炸药爆炸时释放的总能量！因此生命传输机操作人员的那句冷静而平淡的“Energize”背后所蕴含的能量其实是与核爆炸中令天地为之变色的蘑菇云所象征的能量不相上下。这种类型的生命传输机的作用过程——尤其是复合过程——是很难如电视上那样点尘不惊的。当然，如果生命传输机只是在原子或分子层次上对人体进行分解和复合，所涉及的能量就会小得多，大约相当于几十到几百公斤 TNT 炸药爆炸时释放的能量①。一般来说，生命传输机对生命体的分解与复合所涉及的物质层次越低，在分解与复合过程中吸收与释放的能量就越多。

---

① 对于爱思考的朋友来说，这一数值是不需要计算就可以得出的。因为普通 TNT 炸药利用的不是别的，正是爆炸物在原子和分子层次上的结合能（叫做化学能）。因此把人体在这一尺度上分解或复合所涉及的能量大致就等于与人体质量相当的 TNT 炸药所能释放的能量。

我们关于生命传输机的讨论到这里就结束了，与星际旅行中的另一个流行的方案——虫洞——相比，生命传输机在理论可行性方面似乎略显乐观。但我们必须看到，这种乐观性在很大程度上是建立在对生命本质的无知之上的，就像在相对论之前人们可以乐观地认为运动速度在原则上是不受限制的。科学是美丽的，它受益于我们的想象力，又转而为想象力插上新的翅膀。但科学同时也是严谨的，它并不是漫无边际的想象。对生命本质的无知绝不是我们乐观的理由。如果我们真的想要寻求一点乐观的话，也许时间是最好的乐观理由，因为《星际迷航》的故事——确切地说是我所看过的那部分故事——发生在 24 世纪，我们还有 300 年的时间来更好地理解生命，理解物理学。也许到那时我们会更好地理解生命传输机——无论它是可行的还是不可行的。

2003 年 1 月 2 日写于纽约

2014 年 12 月 1 日最新修订

# 虫洞：遥远的天梯

## 一、引言

1985 年的一个学期末，加州理工大学(California Institute of Technology)的理论物理学教授索恩(Kip S. Thorne)刚刚上完一学年的课，正慵懒地靠在办公室的椅子上休息，电话铃声忽然响了起来。打来电话的是他的老朋友，著名行星天文学家萨根(Carl Sagan)。萨根当时正在撰写一部描写人类与外星生命首次接触的科幻小说。写作已近尾声，但身为科学家的萨根希望自己的作品——虽然只是一部科幻小说——尽可能地不与已知的物理学理论相矛盾。在这部小说中，萨根安排女主人公通过黑洞(blackhole)穿越了 26 光年的距离，到达遥远的织女星(Vega)附近。这是整部小说中最具震撼力的情节，但从物理学的角度看，却也是最可疑的细节。于是萨根打电话给从事引力研究的索恩，为这一细节寻求技术咨询。在经过一番思考和粗略

的计算后，索恩告诉萨根：黑洞是无法用做星际旅行的工具的。他建议萨根使用虫洞(wormhole)这一概念，这便有了随后出版，并被拍成电影的著名科幻小说《接触》(*Contact*)。

萨根的小说顺利地出版了，索恩对虫洞的思考却没有因此而结束。

三年后，索恩和他的学生莫里斯(Mike Morris)在《美国物理杂志》(*American Journal of Physics*)上发表了一篇题为《时空中的虫洞及其在星际旅行中的用途》(*Wormhole in spacetime and their use for interstellar travel*)的论文，由此开创了对所谓**可穿越虫洞**(traversable wormhole)进行理论研究的先河①。作为教学性刊物的《美国物理杂志》也因此有幸在一个全新研究领域的开创上留下了值得纪念的一笔。

莫里斯和索恩的文章在虫洞研究中具有奠基性的意义，不过虫洞这一概念却并非他们两人首先提出的。早在1957年，美国物理学家惠勒(John Archibald Wheeler)和学生米斯纳(Charles W. Misner)就在一篇文章中提出了这一概念。那篇文章讨论的主题是所谓的“几何动力学”(geometrodynamics)，那是一种试图把物理学几何化的理论。米斯纳和惠勒的“几何动力学”后来并没有走得很远，但他们在文章中提出的虫洞这一概念却在事隔30多年后得到了全新的发展，并成为了以星际旅行为题材的科幻小说的标准词汇，可谓是“有心栽花花不开，无心插柳柳成荫”。

## 二、什么是虫洞？

那么究竟什么是虫洞呢？形象地说，虫洞是连接两个空间区域的一种“柄”状的结构。图15便是一种很流行的虫洞图示，图中倒U字形曲面代表

① 所谓“可穿越虫洞”，广义地讲，是指允许光信号穿越的虫洞；狭义地讲，则是指允许星际飞船穿越的虫洞。本文所讨论的是后一种。

我们生活在其中的空间，连接两个空间区域 A 和 B 的直线段代表的便是这种“柄”状结构，即虫洞。图 15 是一种抽象化的图示，连接 A 和 B 的直线段实际上代表的是具有一定线度的结构。不难看到，由于这种“柄”状结构的存在，在 A 和 B 之间存在着两种不同类型的路径：一种由曲线表示，代表在普通空间中的路径；另一种由直线段表示，代表由于虫洞的存在而形成的新路径。由图 15 可以看到，沿直线段从 A 到 B 显然要比沿曲线近得多。通常科幻小说——包括前面提到的萨根的小说《接触》——所描述的通过虫洞的星际旅行，就是沿图中直线段进行的。

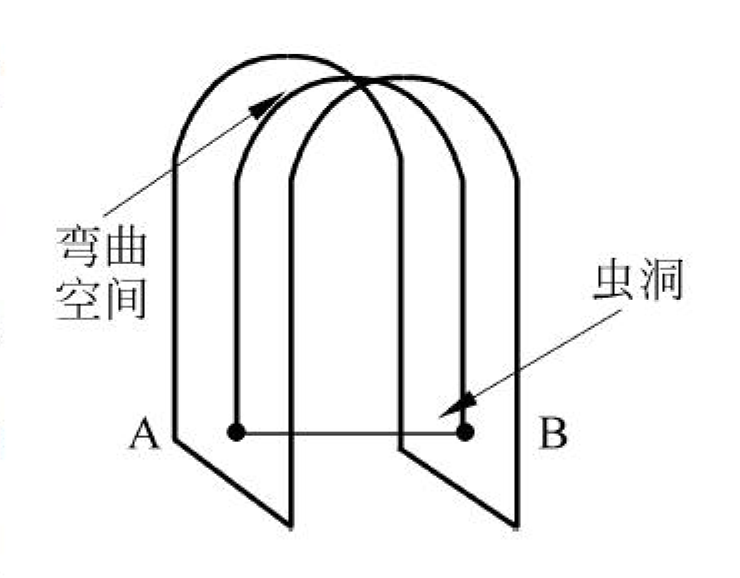


图 15　一种典型的虫洞

在虫洞的研究中，图 15 所示的虫洞被称为“宇宙内虫洞”(intra-universe wormhole)，它连接的是同一个宇宙中两个不同的空间区域。除此之外，在理论上还有一类所谓的“宇宙间虫洞”(inter-universe wormhole)，所连接的是两个不同的宇宙。科幻小说中的虫洞通常属于前一类。不过由于这两类虫洞的差别仅在于空间的大范围拓扑结构，对于讨论虫洞本身的结构来说，它属于哪一类并不重要。

在进一步讨论虫洞之前，我们先来澄清一个或多或少存在于文献中的概念误区（或者说即便在文献作者的心中并无误区，却特别容易在读者之中造成误会的概念），那就是**虫洞的存在并不意味着它们就一定是空间中的捷径**(short-cut)。换句话说，虫洞的存在并不意味着它们就一定能提供一种有意义的星际旅行路径。仔细观察图 15 不难发现，虫洞之所以成为连接 A 和 B 之间的捷径，完全是由于空间弯曲成了倒 U 字形所致。按照广义相对论，空间（确切地说是时空）的弯曲是由物质分布决定的，因而图 15 所表示的虫洞除了虫洞本身外，还对远离虫洞的背景空间中的物质分布作了十分苛刻的假定。如果不做这种相当人为的苛刻假定，虫洞的结构更有可能类似于图 16 所示。在图 16 中，由虫洞所形成的连接 A 和 B

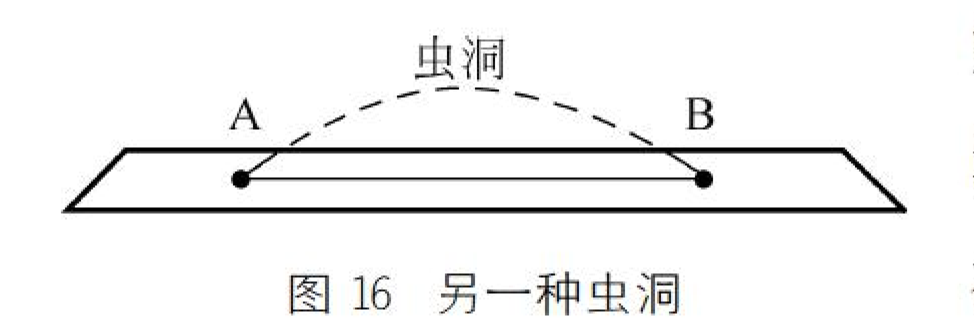


图 16　另一种虫洞

的路径(即虚线路径)要比普通空间中的路径更长。很明显,利用图 16 所示的虫洞进行 A 和 B 之间的星际旅行是很不明智的。因此在概念上,**虫洞并不等同于星际旅行的捷径**。

## 三、萨根式的问题

尽管如此,虫洞无论对于物理学家、天文学家,还是星际旅行家来说,都依然是一个极富魅力的概念。前面提到的行星天文学家萨根对星际旅行所涉及的许多问题有一种很独特的提法,即从一个**无限发达的文明**(infinitely advanced civilization)的角度来看待星际旅行问题的可行性。对于虫洞,一个“萨根式”的问题可以表述为:

一个无限发达的文明是否有可能利用虫洞作为星际旅行的工具?

萨根所谓的“无限发达的文明”指的是在物理规律许可的情况下拥有一切能力的智慧生命。对于这样的智慧生命来说,图 15 和图 16 所示的虫洞或许并无实质区别。只要虫洞存在,即便其结构如图 16 所示,他们或许也有能力通过改变背景空间的曲率使之变为图 15 的形式。因此在这种“萨根式”的问题中,背景空间的具体结构有可能并不重要。

要利用虫洞作为星际旅行的工具当然首先得要有虫洞。宇宙间究竟有没有虫洞呢?这归根结底是一个观测问题。但起码到目前为止的答案是令人失望的,那就是迄今并未发现任何有关虫洞存在的直接或间接证据。因此现阶段我们对虫洞的探讨仅限于理论范畴。自莫里斯和索恩以来,物理学家们在对虫洞的研究上又获得了一些重要结果。这些结果主要是在有关引力和时空的经典理论——广义相对论——的框架内获得的。经过近一个世纪的研究,物理学家们对广义相对论的数学结构已经了解得相当透彻。尤其是自 20 世纪 60 年代以来,随着现代微分几何手段的应用,许多非常普遍的命题被相继证明,其中的一些对于虫洞研究有着十分重要的意义。

为了获得可作为星际旅行工具的虫洞,一个无限发达的文明可作两方面

的努力：

(1) 如果宇宙中不存在虫洞，他们可以试图“创造”虫洞。

(2) 如果宇宙中存在虫洞，他们可以试图“改造”虫洞，使之适合于星际旅行的需要。

下面我们就分头介绍一下这两方面的努力。

## 四、虫洞的“创世记”——恼人的因果律

先来谈谈第一方面的努力，即“创造”虫洞。

所谓“创造”虫洞，指的是在原本没有虫洞的空间区域中产生出虫洞来。我们已经知道，虫洞是空间中的一种“柄”状结构，在拓扑学上具有这种“柄”状结构的空间被称为是复连通的，没有“柄”状结构（即没有虫洞）的普通空间则是单连通的。因此从拓扑学的角度讲，“创造”虫洞意味着使空间的拓扑结构发生变化。

那么空间的拓扑结构有可能发生变化吗？物理学家们对此进行了一系列的研究。1992 年，著名英国理论物理学家霍金（Stephen Hawking）证明了这样一个定理。

［**定理**］ *在广义相对论中，如果空间的拓扑结构在一个有界的区域内发生了变化，那么在这个变化所发生的时空范围内存在闭合类时曲线。*

不熟悉相对论的朋友可能不知道什么叫做“类时曲线”（timelike curve）。在相对论中，类时曲线是**物理上可以实现**的有质量物体在**时空中**的运动轨迹。一个物体在空间中的运动轨迹闭合是十分寻常的事情，比如钟摆的运动，行星的运动，其在空间中的运动轨迹在适当的参照系中都是（近似）闭合的。但一个物理上可以实现的运动在**时空中**的运动轨迹闭合（即形成所谓“闭合类时曲线”）却是非同小可的事情。因为时空中的轨迹不仅记录了运动所经过的所有空间位置，而且还记录了它经过各空间位置的时刻。因此时空轨迹的闭合意味着不仅在空间上回到原点，而且在时间上也回到原点。换句话说，时空轨迹

的闭合意味着时间失去了实际意义上的单向性，或者说构造时间机器成为了可能！

我们都知道，自然万物的演化具有明显的不可逆性，最直接的经验莫过于我们的生命本身，从出生到成长，从衰老到死亡，每一步都不可抗拒、无可逆转。时间的单向性是物理学乃至全部自然科学中最基本的观测事实之一。如果时间不是单向的，那么物理世界中的因果关系也将不复存在，因为一个逆时间而行的旅行者可以在“结果”发生之后返回过去将产生结果的“原因”破坏掉①。

因此霍金所证明的定理可以通俗地表述为：

［**定理（通俗版）**］ 在广义相对论中，“创造”虫洞意味着放弃因果律。

如果放弃因果律，那么不仅物理学的大部分将会被改写，连科学本身的存在都将受到挑战。因为科学本质上就源于人类对自然现象追根溯源的努力，而正是因果律的存在使得这种努力成为可能。因此，依据霍金所证明的上述定理，在有足够证据表明因果律可以被破坏之前，我们必须认为改变空间的拓扑结构（即“创造”虫洞）是被广义相对论所禁止的。

广义相对论是现代物理学中最优美的理论之一，是引力理论和现代时空观念的基石，但它只是一个经典理论。物理学家们普遍认为，对引力和时空的真正描述就像对宇宙中其他基本相互作用的描述一样，必须是量子化的。对广义相对论的量子化被称为量子引力理论。

那么在量子引力理论中情况又如何呢？

早在量子理论出现之初物理学家们就已发现，许多被经典理论所禁止的过程在量子理论中会成为可能，比如电子有可能出现在经典理论不允许出现的区域中。由此带来的一个很自然的问题就是：空间拓扑结构的改变会有幸

① 严格地讲，时间的非单向性（或闭合类时曲线的出现）并不一定导致因果律的破坏。有些物理学家试图通过引进所谓的“自洽性假设”（consistency conjecture）来协调时间的非单向性与因果律之间的矛盾。不过从目前的研究结果来看，这种“自洽性”的一种很有可能的体现方式就是物理规律自动阻止闭合类时曲线的出现。

成为这种量子过程“大家庭”中的一员吗？遗憾的是，对这一问题目前还没有明确答案。引力的量子化是当今理论物理面临的最困难的问题之一，迄今为止不仅尚未建立完整的理论，连一些基本的出发点也还在争议之中。在对量子引力理论的早期研究中，人们曾经设想时空就像海面一样，从大尺度上看平滑如镜，随着尺度的缩小渐渐显出起伏，当尺度缩小到一定程度时，就可以看到汹涌的波涛和飞散的泡沫。这个极小的尺度被称为普朗克尺度（Planck scale）。按照这种设想，在普朗克尺度上时空的结构会出现剧烈的量子涨落，不仅空间的拓扑结构可以发生变化，甚至还会产生所谓的时空泡沫（spacetime foam）。

但是，这种有关量子时空的直观设想在量子引力理论的各个具体方案中均遇到了不同程度的困难。初步的分析表明，量子引力理论并不完全禁止空间拓扑结构的改变，但是**由产生虫洞所导致的空间拓扑结构的改变即使在量子引力理论中也极有可能是被禁止的**。

因此我们可以有保留地认为，就目前我们所了解的物理学规律而言，“创造”虫洞有可能是一件连无限发达的文明也无法做到的事情。

## 五、虫洞工程学——负能量的困惑

接下来谈谈第二方面的努力，即“改造”虫洞，使之适合于星际旅行的需要。

即便“创造”虫洞是不可能的，一个无限发达的文明仍然可以通过改造宇宙中已经存在的虫洞（如果有的话），使之成为可穿越虫洞①。这并不改变空

① 有人也许会问，如果“创造”虫洞是不可能的，那么所谓“已经存在”的虫洞从何而来呢？这是一个很有趣的问题，我们都知道能量守恒是物理学上的一个基本定律，也就是说物质是不能无中生有的，那么宇宙中的物质从何而来呢？这两个问题有相似之处，由于我们对于宇宙本身的由来还知之甚少，因此这些问题都还没有答案。我们把宇宙中“已经存在”虫洞作为这一节的出发点，不仅仅是把它作为一种可能性来看待，同时也是考虑到“创造”虫洞未必真的已被物理定律所严格排除。在这种情况下，假定存在虫洞（不论其来源），考虑如何将之改造并维持为可穿越虫洞是一个不无意义的问题。

间的拓扑结构，从而不违背任何禁止空间拓扑结构改变的物理学定理。

那么，改造一个可穿越虫洞——或者更具现实意义地说，维持一个改造后的可穿越虫洞——需要什么样的条件呢？

前面提到的莫里斯和索恩的文章的主要贡献就是对这一问题进行了定量的分析。他们研究了维持一个稳定的球对称虫洞所需要的物质分布。所谓球对称虫洞，指的是虫洞的出入口——即俗称为“嘴巴”(mouth)的部位——是球对称的。莫里斯和索恩发现，为了维持这样一个虫洞，在虫洞所形成的通道的最窄处——即俗称为“喉咙”(throat)的部位——必须存在负能量的物质(图 17)。莫里斯和索恩的分析虽然对虫洞作了球对称这样一个简化假设，但是运用广义相对论及现代微分几何手段所做的进一步研究表明，他们得出的**维持虫洞需要负能量物质**的结论却是普遍成立的。

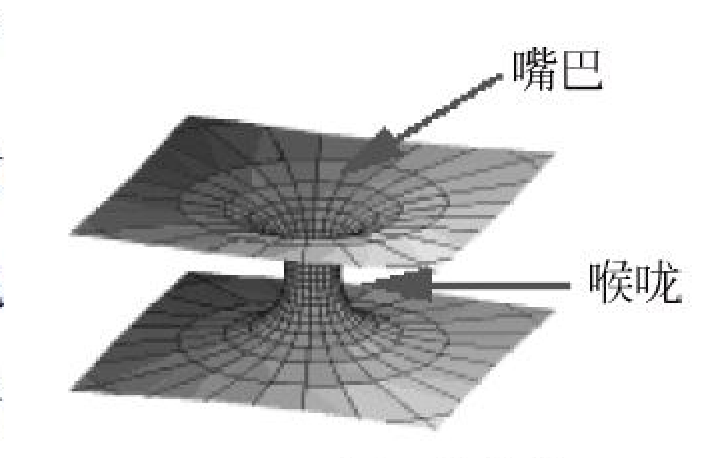


图 17　虫洞的结构

因此，想当一名虫洞工程师，首先得有负能量物质。

那么，什么是负能量物质呢？举一个简单的例子来说，学过牛顿定律的人都知道，用力推一个箱子，箱子就会沿推力的方向运动，推力的大小等于运动的加速度与箱子质量的乘积(假定阻力可以忽略)。这是大家熟悉的结果①。但假如把箱子换成虫洞工程师的负能量箱子，情况就大不相同了。由于负能量箱子的质量小于零，若牛顿定律还能套用的话，加速度与推力的方向就变得彼此相反了。这表明你用力去推一个负能量箱子，非但不能把它推开，箱子反而会朝你滑过来！显然我们谁也没见过这么古怪的箱子，迄今为止人类在宏观世界中发现的所有物质都具有正能量，物质越多，通常能量也就越高。按照定义，只有一无所有的真空的能量才为零，而负能量意味着比一无所有的真空

① 这里所说的质量是“惯性质量”(inertial mass)，另外还有一类所谓的“引力质量”(gravitational mass)。在广义相对论中，这两类质量是相等的。另外在相对论中质量是能量的一种，因此本文对负质量和负能量不作区分。

具有“更少”的物质，这在经典物理学中是近乎于自相矛盾的说法。

但量子理论的发展彻底改变了经典物理学关于真空的观念。在量子理论中，真空不仅具有极为复杂的结构，而且是高度动态的，每时每刻都有大量的虚粒子对产生和湮灭。在这种全新的真空图景下，负能量至少在概念层面上不再是不可思议的了。事实上，早在 1948 年，荷兰物理学家卡西米尔（Hendrik Casimir）就在理论研究中发现真空中两个平行导体板之间会出现负的能量密度，并由此预言了存在于这样一对导体板之间的一种微弱的相互作用。后来人们在实验上定量地证实了这种被称为卡西米尔效应（Casimir effect）的相互作用，从而间接地为负能量的存在提供了证据。20 世纪 70 年代，霍金等物理学家在研究黑洞的辐射效应时发现，在黑洞的事件视界（event horizon）附近也会出现负的能量密度。20 世纪 80 年代，物理学家们又发现了所谓的压缩真空（squeezed vacuum），即量子态分布异常的真空，在这种真空的某些区域中同样会出现负的能量密度。

所有这些令人兴奋的研究都表明，宇宙中看来的确是存在负能量物质的。

但可惜的是，仅仅存在是不够的，还有数量的问题需要考虑。这方面的结果却极不容乐观，因为迄今所知的所有负能量物质都是由量子效应产生的，从而数量极其微小。拿卡西米尔效应来说，计算表明，一对平行导体板之间的负能量所对应的质量密度 $\rho$ 大约为（其中 $\rho$ 以千克每立方米为单位，平行导体板的间距 $d$ 以米为单位）

$$\rho \approx -\frac{10^{-44}}{d^4}$$

这个结果表明如果平行导体板间距为一米的话，所产生的负能量的质量密度只有 $10^{-44}$ 千克每立方米，相当于在每 10 亿亿立方米的体积内才有相当于一个基本粒子质量的负能量物质！

其他量子效应产生的负能量密度也大致相仿，只需把平行导体板间距换成那些效应所涉及的空间尺度即可。由于负能量的密度与空间尺度的四次方成反比，因此在任何宏观尺度上由量子效应产生的负能量都是微乎其微的。

另一方面，物理学家们对维持一个可穿越虫洞所需要的负能量物质的数量 $M$ 也做了估算，结果发现（$M$ 以地球质量为单位，虫洞半径 $R$ 以厘米为单位）：

$$M \approx -R$$

也就是说仅仅为了维持一个半径为一厘米的虫洞[①]，就需要相当于整个地球质量的负能量物质！而且虫洞的半径越大，所需的负能量物质就越多。为了维持一个半径为一千米的虫洞所需的负能量物质的数量竟相当于整个太阳系的质量！

这无疑是一个令所有虫洞工程师头疼的结果。因为一方面，迄今知道的所有产生负能量物质的效应都是量子效应，所产生的负能量物质的数量即使用微观尺度来衡量也是极其微小的。而另一方面，为了维持任何宏观意义上的虫洞所需的负能量物质的数量却是一个天文数字！

## 六、穿越虫洞——张力的挑战

虽然数字看起来不那么乐观，但是别忘了我们是在考虑一个“萨根式”的问题。我们的想象力已经无数次地低估过人类自身科学技术的发展，因此让我们姑且对来自“无限发达的文明”的虫洞工程师的技术水平做一个比较乐观的估计：假定他们利用某种远不为我们所知的技术手段真的获得了相当于整个太阳系质量的负能量物质，并成功地维持住了一个半径为 1 000 米的虫洞。

他们是否就可以利用这样的虫洞进行星际旅行了呢？

初看起来，半径 1 000 米的虫洞似乎应当满足星际旅行的要求了，因为 1 000 米的半径在几何尺度上已经足以让相当规模的星际飞船通过了。看过科幻电影的人可能对星际飞船穿越虫洞的特技处理留有深刻印象。从屏幕上看，飞船穿越的似乎是时空中一条狭小的通道，飞船周围充斥着由来自遥远天

① 这里的半径是指周长除以 $2\pi$。（请读者想一想为什么要作这个注释？）

际的星光和辐射组成的无限绚丽的视觉幻象(图 18)。

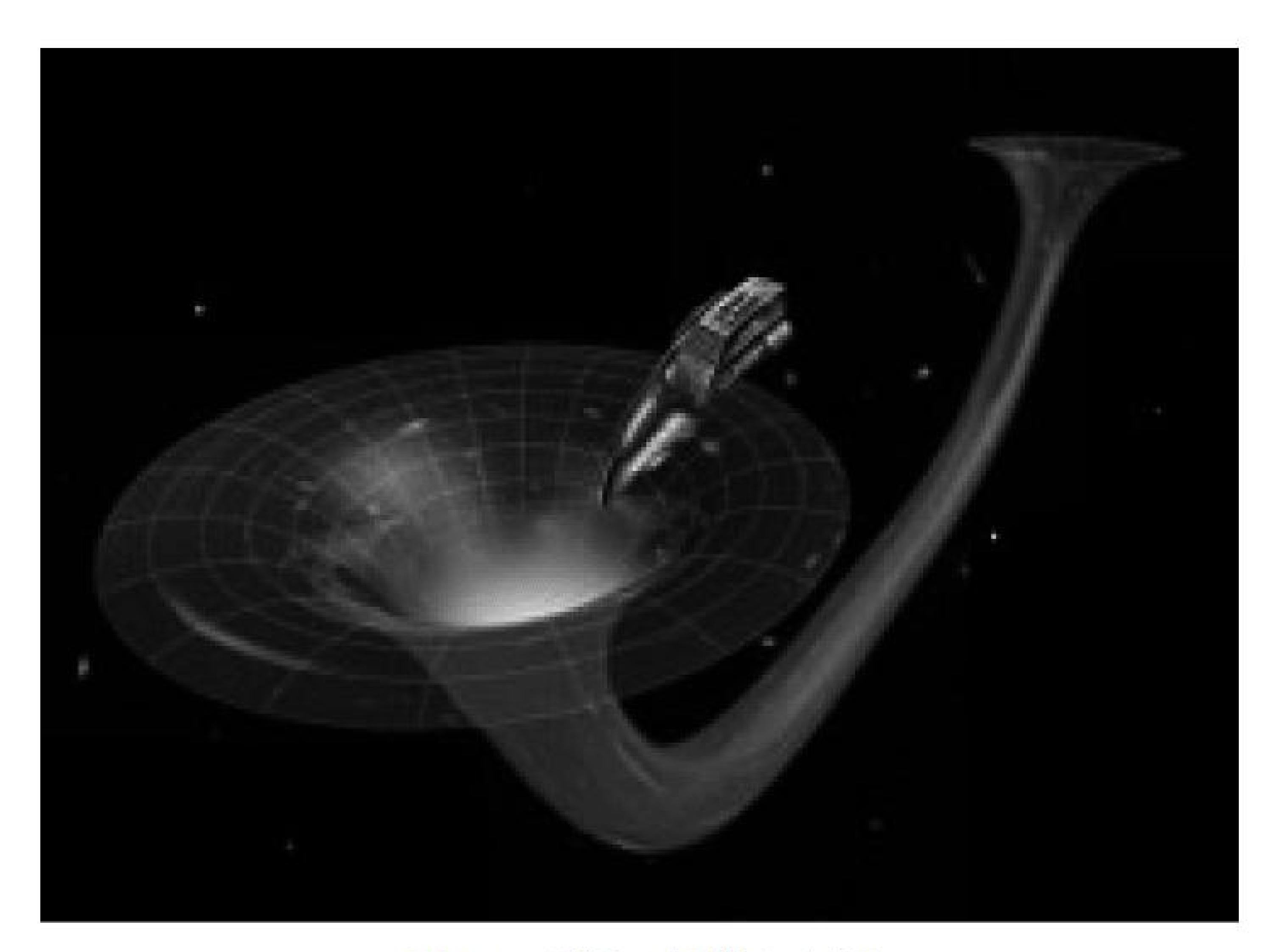


图 18　星际飞船进入虫洞

但实际情况远没有那样诗情画意。

事实上，为了能让飞船及其乘员安全地穿越虫洞，几何半径的大小并不是星际旅行家所要考虑的主要问题。按照广义相对论，为了维持像虫洞那样时空高度弯曲的结构，必须依靠由负能量物质提供的巨大张力。而当飞船及其乘员穿越虫洞，尤其是穿越负能量物质密集的区域——即虫洞的“喉咙”部位——时，将几乎无可避免地会遭遇到这种张力。由于无论飞船还是飞船乘员，他们所能承受的张力都是有限的，因此穿越虫洞时所会遭遇到的张力大小对于星际旅行来说是至关重要的。那么这种张力究竟有多大呢？以球对称的虫洞为例，计算表明，在虫洞的“喉咙”部位，张力的大小约为

张力≈(物质所能承受的最大张力)/(以光年为单位的虫洞半径的平方)

这里“物质所能承受的最大张力”指的是物质中的原子结构所能承受的最大张力。超越了这一极限，连组成物质的原子都将受到破坏，更遑论像飞船或飞船乘员那样的宏观物质了。这恐怕是任何程度的文明——只要他们的生存还离不开物质形体——都很难突破的物理极限。从上述结果中我们看到，穿越虫

洞所会遭遇到的张力大小与虫洞半径的平方成反比，虫洞的半径越大，张力就越小，从而也就越适合于作为星际旅行的通道。特别需要看到的是，**半径小于一光年的球对称虫洞由于穿越时所会遭遇到的张力大小超过物质所能承受张力的理论极限，将很可能无法作为星际旅行的通道。**

虽然以上都是比较粗略的估算，具体数值会因虫洞结构的不同而有所不同。但在数量级的意义上，这种估算已足以使我们看到维持一个可供星际旅行用的虫洞所面临的巨大的“工程学”困难，那就是：一方面，为了能让星际飞船安全通过，虫洞的半径至少要在一光年以上；另一方面，我们在前面已经介绍过，维持一个半径一千米的球对称虫洞所需的负能量物质数量约相当于整个太阳系的质量，且半径越大，所需的负能量物质也越多（与半径成正比），而一光年大约是 10 万亿千米，因此维持一个半径一光年的球对称虫洞所需的负能量物质数量约相当于太阳系质量的 10 万亿倍！

“太阳系质量的 10 万亿倍”是个什么概念呢？我们知道，整个银河系中所有发光星体的总质量大约是太阳系质量的 1 000 亿倍，**因此维持一个可供星际旅行用的最小的球对称虫洞所需的负能量物质数量约相当于银河系中的所有发光星体质量总和的 100 倍！**如果考虑到生物体所能承受的张力要远小于理论极限，对虫洞半径的要求将更高，所需的负能量物质的数量则将比上述估计值更大。使用数量如此惊人的物质，别说这些物质都是迄今尚未在任何宏观尺度上被发现的负能量物质，即便是普通的物质，也是近乎于天方夜谭的想法。

总体来说，目前还不清楚存在于微观尺度上的负能量物质是否有可能积累成宏观数量，如果这种积累是可能的，那么将一个已经存在的虫洞改造并维持成适合星际旅行的虫洞在纯理论上是可能的。但改造并维持那样的虫洞所需的负能量物质的数量即便从宇宙学尺度上看也是极其惊人的。这种数量对于任何存在于我们这个宇宙中的文明——哪怕是无限发达的文明——来说，恐怕都是工程学上一个不可逾越的困难。

## 七、结语——遥远的天梯

在我们即将结束对虫洞的讨论时①，我想起了远古神话中关于“天梯”(ladder to heaven)的一些传说。在远古的年代里，很多人幻想着天空中有一个圣洁而永恒的天堂，人的灵魂能在那里得到永生。虽然谁也不确定天堂离我们有多远，但有些人幻想着存在一些神秘的地方，人们可以从那里攀上天堂，那便是有关“天梯”的传说。古埃及的法老们曾经相信宏伟的金字塔可以成为他们的天梯；藏民们的一种传说，则认为天梯是神山上的一株巨树。从某种意义上讲，虫洞仿佛是一种现代版的“天梯”，一端连着古老而执着的梦想，一端连着遥远而璀璨的星空。

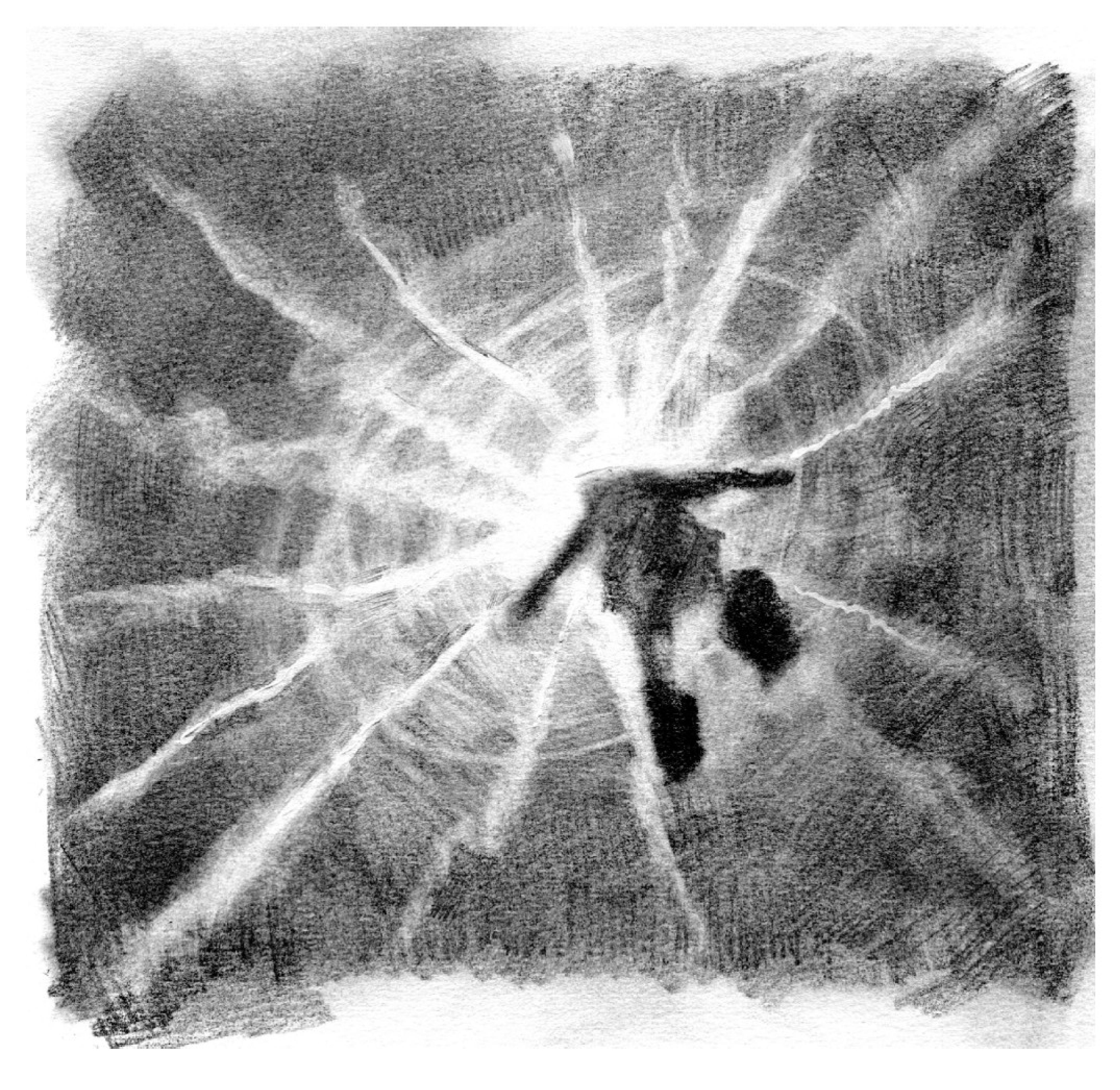
梦想与现实往往是有距离的，任凭虔诚的信徒们千百年不懈地期盼和寻觅，传说中的天梯终究没有被找到。人类对可穿越虫洞的研究才进行了短短十几个年头，下断语还为时过早。但从迄今所得的结果来看，利用虫洞进行星际旅行大致是介于“理论上不可能”和“实际上不可能”之间。在能够想象得到的将来，利用虫洞进行星际旅行很可能就像寻找遥远的天梯一样，只能是一个美丽却难圆的梦。

2002 年 9 月 26 日写于纽约

2014 年 12 月 4 日最新修订

---

① 有关虫洞的深入分析，以及其他一些值得讨论的方面，比如虫洞与时间旅行之间的关系，量子辐射效应对虫洞的作用等，可参阅拙作《从奇点到虫洞：广义相对论专题选讲》(清华大学出版社，2013 年)。



绘画：张京

# 时间旅行：科学还是幻想？[①]

## 一、从《时间机器》讲起

众所周知，迄今为止人类在空间与时间上获得的自由度是很不相同的。我们可以沿空间方向作自由运动，却无法随意驾驭时间。时间就像一条漫漫长河，世间万物仿佛是河里的漂浮物，只能随波逐流。

现实的尽头往往就是幻想的起点。如果时间是一条长河，那么在这长河之中是否能有船只呢？漂浮物只能随波逐流，船只却可以劈波斩浪。如果时间长河中能有船只，我们就可以乘坐这种船只进行时间旅行，既可以窥视未来，也可以重返往昔，说不定还能改变历史。在科幻小说中，这种假想的船只被称为“时间机器”。

有关时间机器最早、最著名的小说是英国科幻作家威尔斯（H. G. Wells）的《时间机器》（*The Time Machine*），发表于 1895 年。不过，威尔斯并不是最早触及时间旅行这一题材的作家，在他之前已经有许多作家涉足过这

① 本文曾发表于《科幻世界》2006 年第 7 期（科幻世界杂志社出版）。

一题材,其中甚至包括美国讽刺小说家马克·吐温(Mark Twain),他发表于1889年的《康州美国佬在亚瑟王朝》(*A Connecticut Yankee in King Arthur's Court*)据说是最早涉及逆向时间旅行的小说。但在那些比威尔斯更早的文学作品中,普遍没有使用像时间机器这样一种可以让人选择"目的地"(确切地讲是"目的时间")的旅行器,并且也极少对时间旅行的机制作哪怕只是科幻意义上的说明。而威尔斯的《时间机器》在这两方面都是突破性的,它很快引起了读者们的巨大兴趣,并于1960及2002年两度被拍成电影,英国甚至为《时间机器》出版100周年发行过纪念邮票。

威尔斯写作《时间机器》的时候,爱因斯坦(Albert Einstein)的相对论尚未被提出,人们对时空的理解大体上还停留在牛顿(Isaac Newton)的绝对时空观上①。但威尔斯却在《时间机器》一书中令人吃惊地提出了将时间作为第四维的观点,与十年后到来的相对论时空观作了戏剧性的遥相呼应。

威尔斯将时间视为第四维,目的是要通过将时间与空间类比来为时间旅行开绿灯。那么现代物理学认可这个绿灯吗?这就是本文所要讨论的内容。

## 二、面向未来与重返过去

我们知道,在牛顿的绝对时空观里,时间和空间不受任何物质及运动的影响(这是"绝对"的主要含义所在)。很明显,在这样的时空观里,时间旅行不具有理论基础,它的存在只是一种幻想。但是狭义相对论的提出对时空观产生了一次重大变革。在狭义相对论中,时间和空间不再是绝对的概念,而是与参照系的选择密切相关。特别是,在运动参照系中时间的流逝会变慢,这是著名的时间延缓效应,它的存在已经被大量物理实验所证实。狭义相对论所带来

① 在1892至1895年间,荷兰物理学家洛伦兹(Hendrik Lorentz)等人曾在研究电磁理论时提出过一些有别于绝对时空观的假设,但这些假设并未成为主流,后来则被相对论所取代。

的这种新结果，为时间旅行开启了第一种具有理论依据的可能性：那就是面向未来的时间旅行成为了可能。

按照狭义相对论，如果有人想要到未来去旅行，他所需要的时间机器就是一艘能以接近光速的高速度运行的飞船。想要到达的未来越遥远，飞船所需达到的速度就越高。如果他想在20年（飞船上的时间）的飞行之后到达两万年（地球上的时间）后的地球上，他所要做的就是让飞船以相当于光速99.99995%的速度飞行10年，然后以相同的速度往回飞。那么20年后，当他回到地球上时，地球上的日历已经翻过了整整两万年，他可以如愿以偿地看到两万年后的人类社会（如果那时候人类社会还存在的话）。可以想象，这样一位来自远古的旅行家将会受到未来的历史学家和考古学家们何等热烈的欢迎。

事实上，不仅未来的历史学家和考古学家将会非常欢迎这样的时间旅行家，与这位时间旅行家同时代的人又何尝不希望他能把自己看到的未来世界的情形带回给大家呢？可惜的是，狭义相对论为面向未来的时间旅行开启了大门，却没能为重返过去的时间旅行提供同样的理论可行性。如果一定要对狭义相对论的数学框架做广义诠释的话，那么只有超光速的运动才可能导致某一类参照系中的时序被颠倒。但是狭义相对论本身在亚光速与超光速之间设置了一个光速壁垒，没有任何已知的物理过程能够使原本亚光速运动的物体——包括人——进入超光速运动状态。因此在狭义相对论的理论框架内，时间旅行家可以到达未来，但却不能重返过去，这与我们在空间中自由自在的运动相比，显然是差得很远的。而且，面向未来的时间旅行不一定需要时间机器才能做到，通过将旅行者冷冻若干年再解冻的手段也可以达到同样的目的。因此时间机器如果存在的话，它真正独特的价值不在于面向未来，而在于重返过去。

那么重返过去的路在哪里呢？

在狭义相对论之后又过了10年，爱因斯坦提出了广义相对论。在广义相对论中，时间和空间不仅如狭义相对论中一样与参照系的选择密切相关，而且

还有赖于物质的分布和运动。由此产生的一个不同于狭义相对论的重要结果是:我们对“未来”的定义不再是绝对的了,它会受到物质运动的影响。在不同时刻、不同地点,“未来”有可能指向不同的方向。这是一个奇妙的结果,它表明时空在某种意义上就像流体一样会受到物质运动的拖曳,甚至连时间的方向都有可能因拖曳而改变。

既然时间的方向可以被物质的运动所拖曳,那么有没有可能存在某种物质的分布与运动,它对时间方向的拖曳如此显著,以至于把未来方向拖曳成过去方向,甚至让不同的时间方向首尾相接,连成一条闭合曲线呢?这样的闭合曲线如果存在,无疑就是一种时间机器。因为沿这种曲线运动的飞船每时每刻都在做正常的飞行,感受到正向的时间流逝,但它的轨迹却不仅在空间上,而且会在时间上回到出发点。如果你乘坐飞船沿这样的曲线做一次为期10年的旅行①,那么在旅行结束时你不仅会回到飞船出发的地方,并且会遇见10年前整装待发的自己②!物理学家们把这种奇妙的曲线称为“闭合类时曲线”,它是时间机器这一科幻术语在广义相对论中的代名词。倘若存在闭合类时曲线,时间旅行就有了理论上的可能性。

那么在广义相对论中,是否存在闭合类时曲线?或者确切地说,是否存在使闭合类时曲线成为可能的物质分布与运动呢?对这个问题,物理学家们做了许多研究。

## 三、广义相对论与时间旅行

1949年,著名逻辑学家哥德尔(Kurt Gödel)在广义相对论中发现了一个非常奇特的解,描述一个如今被称为“哥德尔宇宙”(Gödel universe)的整体旋

① 这里“为期10年”指的是飞船上的时间。

② 事实上,不仅旅行结束时的你会看到10年前的自己,10年前的你在出发时也会看到10年后凯旋归来的自己。假如你在出发时什么都没看到,说明旅程中必定会发生意外,使你无法回到旅行的起点。在这种情况下,你或许应该取消旅行!

转的宇宙。在这种宇宙中，物质的旋转对时间方向会产生拖曳作用，离旋转中心越远，拖曳作用就越显著。在足够远的地方，拖曳作用足以形成闭合类时曲线。因此，在哥德尔宇宙中只要让飞船沿某些远离旋转中心的轨道运动，原则上就可以实现时间旅行。哥德尔这位曾经以哥德尔不完全性定理（Gödel's incompleteness theorems）震撼整个数学界的逻辑学家，又用他的旋转宇宙震动了包括爱因斯坦本人在内的许多物理学家。

可惜的是，哥德尔宇宙并不符合天文观测。首先，我们所生活的宇宙并不存在整体的旋转[①]；其次，在哥德尔宇宙中宇宙学常数是负的，而我们观测到的宇宙学常数却是正的。因此我们所生活的宇宙显然不是哥德尔宇宙。不仅如此，定量的计算还表明，即便我们真的生活在一个哥德尔宇宙中，也很难实现时间旅行，因为沿哥德尔宇宙中的闭合类时曲线运行一周所需的时间与宇宙的物质密度有关，对于我们所观测到的物质密度而言，沿闭合类时曲线运行一周起码需要几百亿年的时间。因此哥德尔宇宙对于时间旅行并无现实意义。

不过，哥德尔宇宙虽然没有现实意义，但它的发现表明广义相对论的确允许闭合类时曲线的存在，这本身就是一个鼓舞人心的结果。自那以后，物理学家们在广义相对论中又陆续发现了其他一些允许闭合类时曲线的解。比如1974年，美国图兰大学（Tulane University）的物理学家梯普勒（Frank J. Tipler）研究了一个无限长的旋转柱体外部的时空[②]，结果发现只要旋转速度足够快，这样的柱体对外部时空所起的拖曳作用也足以形成闭合类时曲线。又比如1991年，普林斯顿大学的天体物理学家高特（John Richard Gott Ⅲ）

① 当然，这是指在现有的观测精度内没有发现宇宙的整体旋转。另外，有读者可能会问：什么是宇宙的整体旋转？这种旋转是相对于什么来定义的？这类问题可以视为是跟奥地利哲学家马赫（Ernst Mach）的观点，即旋转必须是相对的，一脉相承。不过，尽管爱因斯坦本人曾经推崇过马赫，但广义相对论事实上并不严格遵循马赫的哲学观点。

② 梯普勒并不是最早研究这一时空的物理学家，早在1937年，荷兰物理学家范斯托克姆（Willem Jacob van Stockum）就曾研究过这一时空，只不过没有像梯普勒那样对其因果特性进行分析。

发现两条无限长的平行宇宙弦以接近光速的速度彼此擦身而过时,也会在周围形成闭合类时曲线。与梯普勒人为引进的旋转柱体不同的是,宇宙弦的存在虽然还没有明确的实验证据,但它是许多前沿物理理论所预言的东西。因此高特的结果可以算是把时间机器在理论上的可能性又推进了一步。

但是梯普勒与高特为了数学上的便利都引进了无限长的物质分布(即“无限长的旋转柱体”和“无限长的平行宇宙弦”),这在现实世界中显然是不可能严格实现的。假如物质的分布不是无限的,还可以得到类似的结果吗?物理学家们对此也做了研究,但情况不容乐观:1992 年,著名物理学家霍金(Stephen Hawking)给出了一个令人沮丧的结果,那就是如果能量密度处处非负,那么试图在任何有限时空区域内建造时间机器的努力要想成功,都必须产生物理学家们最不想看到的东西——时空奇点①。时空奇点对于研究广义相对论的人来说是并不陌生的,它具有一系列令人头疼的性质,比如物质的密度发散,时空的曲率发散,等等②。虽然没有人确切知道时空奇点的出现会对时间旅行产生什么影响,但这种影响很可能是凶多吉少的。

霍金的这个结果对于建造时间机器无疑是坏消息,但细心的读者也许注意到了,这个结果中有一个限制条件,那就是“能量密度处处非负”。这个条件粗看起来是非常合理的,但我们在介绍虫洞的时候已经提到过,负能量物质的存在不仅在理论上是可能的,而且已经得到了实验的证实。

既然负能量物质可以存在,那么霍金的结果(确切地说是其中的结论部分)就有可能被避免。这方面的研究事实上早在霍金的结果出现之前就已经有人进行了——当然目的不是为了避免当时尚未出现的霍金的结果:加州理工大学的物理学家索恩(Kip Thorne)与学生莫里斯(Mike Morris)等人在1988 年发表的一项有关“可穿越虫洞”(traversable wormhole)的研究中,发现

---

① 确切地讲,许多物理学家都得到过类似的结果,霍金的只是其中之一。

② 奇点的严格定义本身就是广义相对论中一个非常棘手的课题,这里叙述的只是某一类奇点的特性,更详细的叙述可参阅拙作《从奇点到虫洞:广义相对论专题选讲》(清华大学出版社,2013 年)。

虫洞不仅是空间旅行的通道，而且还可以作为时间旅行的工具——只要让虫洞的出入口以接近光速的速度作适当的运动，就可以将虫洞转变成时间机器①。由于虫洞中含有负能量物质，因此他们这种时间机器可以避免霍金的结果，不导致时空奇点（从这个意义上讲，负能量物质还真是很有“正能量”）。索恩等人的这一研究把科幻小说中最具魅力的两个概念——虫洞与时间机器——联系在了一起，集“万千宠爱”于一身，很快就成为了建造时间机器的热门方案。

但是，索恩等人的虫洞时间机器虽然可以避免霍金的结果，却立即遇到了另一个棘手的问题，那就是虫洞一旦成为时间机器，在类时曲线闭合的一刹那，任何微小的量子涨落都有可能通过那样的虫洞返回过去，与它本身相叠加。这种叠加过程可以在零时间内重复无穷多次，由此产生的自激效应足以在瞬间将时间机器彻底摧毁！这种效应不仅危及索恩等人的“虫洞时间机器”，对其他类型的时间机器也同样具有威胁。1992 年，霍金干脆提出了著名的时序保护假设（chronology protection conjecture），认为自然定律不会允许建造时间机器。不过迄今为止，这还只是一个假设，而且霍金的论据也不是无懈可击的，对时间机器的理论可行性持乐观看法的物理学家们陆续提出了一些模型来突破霍金对时间机器的封杀。这方面的讨论目前仍在继续。

## 四、时间旅行与因果佯谬

有关时间机器的讨论除了探讨它的理论可行性外，还有一个非常重要的方面，那就是探讨时间机器假如存在，我们能用它来做什么？

粗看起来，这似乎不成之为问题，既然能够做时间旅行，那么到达目的时

① 具体地说，让虫洞成为时间机器所需的最简单的运动是那种使虫洞两个出口之间的外部空间距离迅速改变，而虫洞本身的长度却不改变的运动。产生这种运动并不容易，但在原则上是可以做到的。关于“虫洞时间机器”的更详细介绍，可参阅拙作《从奇点到虫洞：广义相对论专题选讲》（清华大学出版社，2013 年）。

间之后自然应该是想做什么就可以做什么——只要不违反物理学定律。但细想一下,事情又不那么简单。举个例子来说,倘若时间旅行者回到自己出生之前,他能够阻止自己父母的相识吗?这似乎不需要违反任何物理学定律。比如时间旅行者若在自己的父母相识之前,向后来会成为自己父亲的那个人开枪,子弹似乎完全可以在不违反任何物理学定律的情况下击中目标,造成致命伤害。但如果那样的行动成功了,我们就会立刻陷入所谓的"因果佯谬"(causality paradox)之中。因为如果时间旅行者的父母因为他的阻挠而没有相识,那么世上就不会有他;而世上如果没有他,他又如何能够返回过去并阻止自己父母的相识呢?像这样的佯谬在考虑时间旅行时数不胜数,它们都起源于时间旅行对因果时序可能造成的破坏。

这类佯谬该如何解决呢?在科幻小说或电影中,解决的方式往往是通过各种巧合。比如前面提到过的威尔斯的《时间机器》在 2002 年被拍摄成影片时,或许是为了对主人公建造时间机器的动机做出某种说明,导演增添了主人公情人被害,他试图重返过去加以挽救的情节。在那段情节中,主人公想尽办法,却总是顾此失彼,他的情人总会以这样或那样的方式死去。显然,同样的手法也可以用来避免时间旅行者阻止自己的父母相识。比方说当时间旅行者正要采取某种手段阻止父母相识时,不小心踩到一块香蕉皮摔伤住进医院,从而错过了时机①。这样的解决佯谬的方式被一些物理学家戏称为"香蕉皮机制"(banana peel mechanism)。在"香蕉皮机制"下,时间旅行者看似能够自由行事,但每当其行为将要导致因果佯谬时,总会受到某些看似偶然的因素干扰,致使行为失败。

---

① 当然,这只是最简单的巧合(不过"香蕉皮机制"因之而命名,故特意举出)。为了情节的需要,我们还可以设想更为复杂的巧合。比方说时间旅行者试图向后来会成为他父亲的那个人开枪,却因为心情矛盾导致枪法失准,没有击中"父亲",却击中了"父亲"的情敌!他试图阻止父母相识的行动非但没有达到目的,反倒为他父母的结合铺平了道路。他的行动不仅没有破坏因果关系,反而成为了维护因果关系所必需的,等等。像这种近乎宿命的巧合在科幻故事中用得也很多。

这种“香蕉皮机制”很适合编写戏剧性的故事情节。但从物理学的角度讲，很难想象物理学定律需要通过如此离奇巧合的方式来解决佯谬①。更何况，香蕉皮机制还有一个致命弱点，那就是它往往只着眼于保证一两个核心事件——比如影片《时间机器》中主人公情人的死亡，或者我们所举的例子中时间旅行者父母的相识——的发生不会被时间旅行所改变，却无法兼顾其他事件。比如影片《时间机器》中主人公的情人以不同方式死亡会在当地报纸上留下不同的报道；我们所举的例子中时间旅行者的摔伤住院也会在当地医院中留下相应的记录。这些事件对特定的故事来说并不突出，但从维护因果时序或历史的角度讲却与核心事件有着同等的重要性。事实上，自然界的各种事件之间存在着千丝万缕的联系，任何看似微小的变化，都有可能通过这种联系逐渐演变成重大事件，这一点对混沌理论中的蝴蝶效应（butterfly effect）有所了解的读者想必不会陌生②。

除香蕉皮机制外，在一些科幻故事中还可以看到另外一种观点，那就是在一定程度上放弃因果律，以扩大时间旅行者的行动自由。在这种观点下，历史可以近乎随意地被改变，并且改变的结果可以影响到现实世界中的许多事情。科幻影片《频率》（*Frequency*）体现的就是这种观点。在那部影片中，主人公虽然没有直接进行时间旅行，但他通过与30年前去世的父亲建立联络，具备了

---

① 尽管如此，还是有物理学家做过这方面的考虑。比如俄国物理学家诺维科夫（Igor Novikov）曾经提出过一个假设，认为物理学定律会——哪怕通过离奇巧合的方式——自动保证不出现因果佯谬。这个假设被称为“诺维科夫自洽性假设”（Novikov consistency conjecture），它可以算是香蕉皮机制的理论版本。不过这个假设一直缺乏具体的实现方式。

② 举个例子来说，如果时间旅行者回到过去后把一块小石头放在路上，然后离开。这样的事件无疑是非常微不足道的，但它有可能导致某位行人因踩到石头而扭伤脚。而这位倒霉的行人有可能恰好是一位物理学家，他正要去做一个有关时间旅行的学术报告，却因为扭伤了脚而取消报告。而那个学术报告的听众中有可能恰好有一位年轻人因为这个报告的影响而投身于时间旅行的研究，并最终成为时间机器的建造者。在这种情况下，时间旅行者放在路上的小石头对历史的影响就扩大成了尖锐的佯谬。因为正是这块石头的出现，使得一位物理学家取消了学术报告，既而又使得一位年轻人因没有听到这个学术报告而不再以时间旅行作为自己的研究方向，而这最终导致了人类没能研制出时间机器。但如果人类没能研制出时间机器，时间旅行者又如何能够放置那块小石头呢？

间接改变历史的能力。在影片中,历史事件的每一次改变都会直接改变 30 年后的现实世界。比如由于历史事件的改变导致主人公母亲意外死亡,30 年后主人公母亲的相片就会从相框中突然消失。显然,这种观点几乎等于放弃已知的物理学定律,比试图保护现实的香蕉皮机制更为离奇。

## 五、凝固长河与平行宇宙

像“香蕉皮机制”或放弃因果律这样的做法,虽然也有物理学家表述过,但总体来说,它们与现实物理学定律之间的差距太大,很少有物理学家会在没有足够证据的情况下,对物理学定律做如此剧烈的变动。对物理学家们来说,更感兴趣的问题是:在现有物理学定律的基础上,能否理解或避免由时间旅行所可能导致的因果佯谬?

对于这一问题,物理学家们尚未形成一致的看法。我们在这里向读者介绍两种主要的观点。

第一种观点认为时间和空间是对物理事件的完整标识。因此一旦时间和空间同时确定,物理事件也就完全确定了。从这个意义上讲,如果我们把时间比作一条长河,那它其实是一条凝固的长河,它的每个截面——对应于一个确定时刻所有物理事件的全体——都是固定的,就像电影胶片一样。按照这种观点,历史只能有一个版本,如果时间旅行者能够回到过去,唯一的可能是他原本就存在于过去。这话听起来有点玄妙,用平直一点的话说就是时间旅行者回到过去后所做的一切都只能精确地演绎历史上已经存在过的一个人。如果他试图阻止自己父母相识,却不小心踩到香蕉皮摔伤住了院,那么在历史上就的确存在过这样一个人,乘坐奇怪的机器从天而降,很不幸地踩到香蕉皮摔伤住了院,伤愈后又乘坐奇怪的机器离去。换句话说,时间旅行者并不能对历史做分毫的改变,他甚至连历史的旁观者都不是,因为他原本就是历史的一部分。这种观点对于热衷时间旅行的人来说无疑是令人失望的,因为如果一切都是不可改变的,那么时间旅行也就失去了最重要的价值。

幸运的是，第二种看待时间旅行的观点要开放得多，这种观点来源于美国物理学家艾弗里特（Hugh Everett Ⅲ）1957 年提出的一种奇特的量子力学诠释——多世界诠释（many world interpretation）[①]。我们知道，量子力学的一个重要特点就是对量子体系进行测量的结果往往是不唯一的。那么，一个具体的测量结果究竟是如何产生的呢？物理学家们提出了许多不同的观点。有些物理学家认为当我们对量子体系做测量时，体系的状态会发生坍缩，我们观测到的测量结果是一个坍缩后的状态。在这种观点中，状态的坍缩是一个不可预测的过程。与之相反，艾弗里特等人的多世界诠释则认为，并不存在这种不可预测的状态坍缩，量子测量的结果是世界分裂为一组平行宇宙。所有量子力学中可能出现的测量结果都是真实存在的，只不过它们分别存在于各自的平行宇宙而非单一世界中。观测者所得到的测量结果，只不过是他（她）所在的平行宇宙中的特定结果而已[②]。如果我们把这种观点运用到时间旅行中，认为时间旅行者不仅跨越时间，而且还跨越不同的平行宇宙，那么所有的佯谬就都迎刃而解了[③]。比如时间旅行者阻止自己父母的相识就不再成为佯谬，因为所有这一切都发生在一个不同的平行宇宙中。在那个宇宙中他的父母原本就不相识，他自己也原本就不曾出生过。这与阻止父母相识的时间旅行者本人出现在那个宇宙中并不矛盾，因为时间旅行者是来自于另一个平行宇宙的，在那个平行宇宙中他父母依然相识。在这种观点下，每个平行宇宙的历史仍然是唯一的，但是所有物理定律许可的历史都会在某个平行宇宙中得

① 艾弗里特是多世界诠释的提出者，不过“多世界诠释”这一术语却是美国物理学家德惠特（Bryce DeWitt）提出的。

② 需要指出的是，多世界诠释的原始表述其实并不依赖于像“多世界”或“平行宇宙”那样的概念。后来流行的“多世界”或“平行宇宙”概念从某种意义上讲是对多世界诠释本身的诠释。

③ 当然，这里所谓的“迎刃而解”，是建立在有着极大争议性的平行宇宙概念之上的，因而本身也是有着极大争议性的。此外，所谓“迎刃而解”，首先还假定所讨论的问题有意义，这同样有可能是不成立的，因为时间旅行完全有可能是如霍金猜测的那样被物理学定律所禁止的，由时间旅行所导致的因果佯谬也因此完全有可能是伪问题。

以实现,时间旅行者虽然无法改变任何一个平行宇宙的历史,却可以自由地选择进入哪一个平行宇宙,他不能改变历史,却可以选择历史①。

## 六、幻想与历史

经过了这些讨论,现在让我们回到本文的标题上来,时间旅行究竟是科学还是幻想?据说索恩与学生发表有关虫洞及时间旅行的论文时,曾经担心被同事们认为是不务正业。但我们在本文中已经看到,在时间旅行这个主题背后有着一系列值得深入研究的物理学课题。事实上,现在的确有一小部分物理学家——其中包括世界顶尖大学的教授——在对这些课题进行认真的研究。这种研究除了试图探讨科幻小说中这些迷人话题的理论可行性外,一个很重要的动机是要探索现有物理学定律的边界,探索在最离奇的情形下物理学定律可以告诉我们什么。从这个意义上讲,时间旅行无疑是一个有着丰富科学内涵的课题。

但是另一方面,从现实可行性上来讲,起码就我们目前所知的物理学定律而言,时间旅行很可能只是一种幻想。我们在前面讨论过许多有可能形成闭合类时曲线的理论模型,撇开它们面临的种种理论难题不论,在那些讨论中我们还忽略了一个很重要的方面,那就是虽然从结构上讲,闭合类时曲线与能让人类使用的时间机器完全类似,但在规模上却有着巨大差异。以索恩等人的虫洞时间机器来说,为了让人类能够使用这种时间机器,虫洞必须是可穿越虫洞。而我们在有关虫洞的介绍中已经看到,建造可穿越虫洞是一件几乎不可能做到的事情,更遑论让虫洞的出入口以接近光速的速度作特定的运动了。因此,索恩的虫洞时间机器无论在理论上是否可能,在现实世界中实现的可能

① 即便按照这种观点,科幻小说中的许多情节也是不可能实现的。比如通过时间旅行者对某个历史事件的干预来改变人类命运就是不可能的。时间旅行者的努力,只能使他自己进入一个人类命运截然不同的平行宇宙中去,而试图通过这一努力来改变自己命运的原平行宇宙中的其他人的命运,将不会因此而改变。

性都是微乎其微的。

限于篇幅，我们有关时间旅行的介绍到这里就告一段落了。十多年前，霍金曾经问过这样一个问题：假如时间旅行是可能的，为什么在我们周围至今尚未充斥着来自未来世界的时间旅行者呢？这个问题的潜台词是：时间旅行者没有来到我们周围，最有可能的原因是时间旅行在整个时间长河中——也就是永远——都没有实现过。当然，霍金并没有把这样的问题当作是对时间机器的一个认真的理论诘难。不过，他的这个问题还是引起了一些物理学家的思考，并且他们找到了一种可能的回答：即我们目前所知的有可能实现时间旅行的理论模型，有一个很可能具有普适性的共同特点，那就是不允许时间旅行者回到时间机器存在之前的年代。因此，假如公元 2500 年有人建造出了时间机器，那么时间旅行者只能访问公元 2500 年之后的年代[①]，他们永远无法来到我们周围，更无法像一些科幻小说描绘的那样，回到史前时代去捕捉恐龙——那些历史已经或将要无可挽回地被时间长河所吞没，就像美国物理学家格林（Brian Greene）所说的：在时间机器建造成功之前的每一个年代，都将成为我们以及我们的子孙后代永远无法触及的历史。

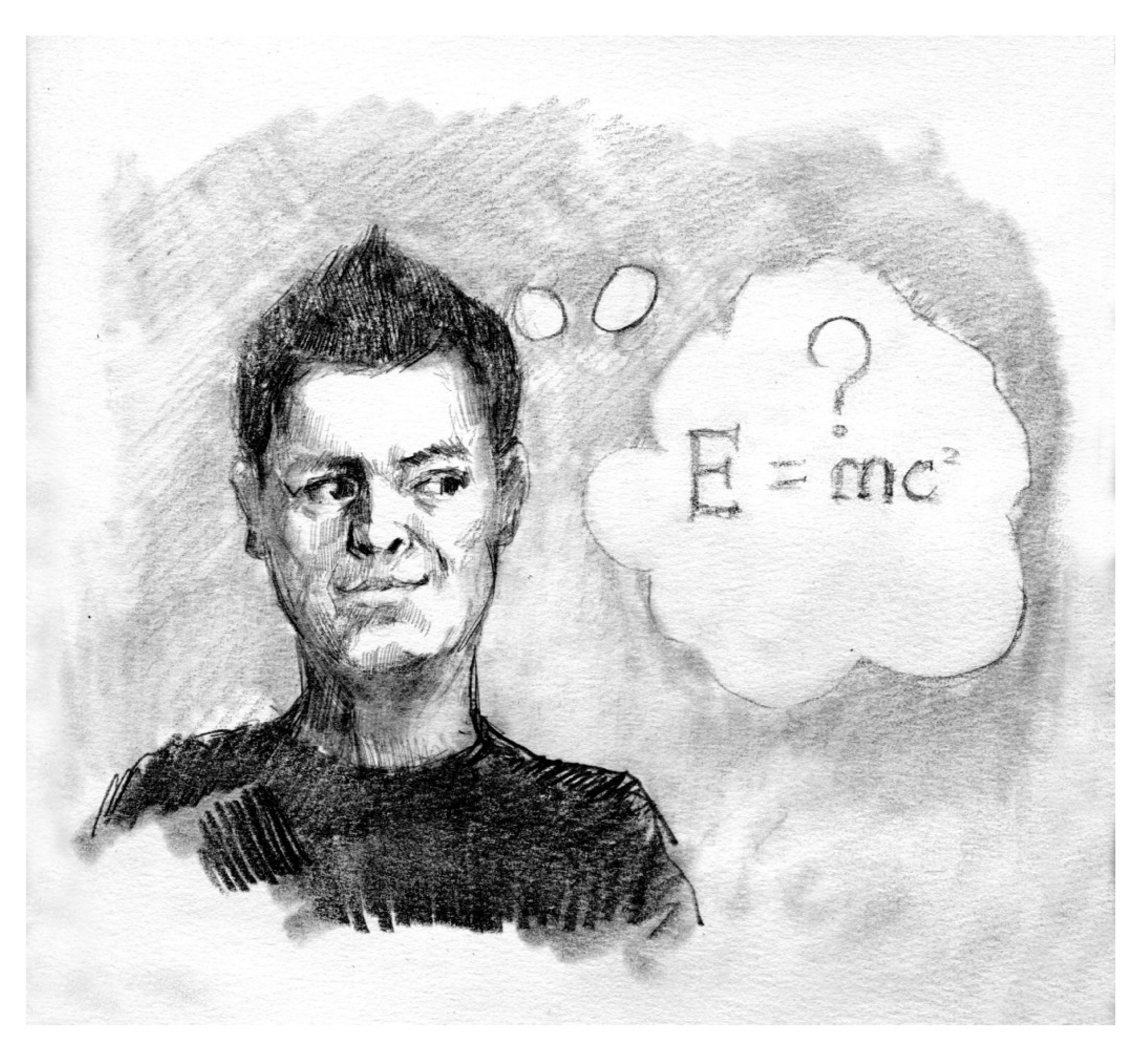
从这个意义上讲，如果时间旅行是可能的话，早一天建造出时间机器就是多拯救一天历史。

2006 年 5 月 18 日写于纽约

2014 年 12 月 7 日最新修订

① 注意，这并不是说时间旅行者只能作面向未来的时间旅行。在时间机器存在之后的那些年代之间，他们的旅行既可以面向未来也可以面向过去，他们只是无法回到时间机器建造之前的年代去。

# 第四部分　其　　他




绘画：张京

# 从民间“科学家”看科普的局限性

半年多前，我在网上偶然发现了一个名为“超弦学友论坛”的网站。

那是一个以讨论超弦理论及相关话题为主的中文学术论坛，设有一个主论坛和一个灌水区，后者是留给与学术无关的话题的。常言道：“林子大了，什么鸟都有”，建一个灌水区可以让不做学术的鸟儿也有个试嗓子的地方。与其他论坛相比，“超弦学友论坛”的最大特点，是有几位中科院及中国科技大学的教授主持，因此秩序相对好些。我初次光顾该论坛的时候，主论坛上有教授和同学们的许多讨论，就像一个网络课堂。但不久前旧地重游，却发现“林子”里的光景已经大变，主论坛上有大批民间“科学家”往来穿梭，在灌水区却发现了一位原先很活跃的教授的踪迹。教授在那里发了一个短短的跟帖，所跟的是他本人被别人转过来的一篇文章。教授在跟帖中写道：

> 谢谢转帖，但我希望尽量不要将我的东西转贴到隔壁，因为隔壁演变成了一个民间“科学”论坛。

这一跟帖对论坛无疑是一个警讯，不久之后论坛的管理员出来删除了一些帖子。

“超弦学友论坛”所遭遇的这种情况在网上是有一定代表性的。互联网的发展给原本需要自费印刷资料、自费前往学校或科研院所推销“理论”的民间“科学家”们提供了极大的便利，使他们亮相的成本大幅降低，“出镜率”也因此大幅提高。大众对民间“科学家”的态度遂成为近年来较有争议的一个话题。

民间“科学家”这一概念并没有一个很严格的定义，因为这是一个具有相当复杂性的群体。往上了看，一部分科学家在其童年或少年时期的思维形式与某些民间“科学家”也有一定的相似性；往下了看，许多伪科学或反科学人士的思维形式与民间“科学家”同样有一定的相似性。粗略地讲，民间“科学家”主要有这样两条特征：

一、民间“科学家”没有接受过系统的科学训练

这一条几乎是定义性的。多数民间“科学家”自己也坦承这一条，就像在过去某个年代里，大家并不避讳自己的赤贫家境一样。这里所说的系统的科学训练并不单单指的是科班出身，完全也可以是达到同等层次的高水平的自学。此外，这里所说的系统的科学训练是以真正学到手为判据的，而不是仅仅混到一个文凭。

二、民间“科学家”无意接受系统的科学训练

这一条往往被人忽略，不过我觉得这一条其实很关键。因为即使是最优秀的科学家，也并非是生来就接受过系统的科学训练的，因此“没有接受过系统的科学训练”并不是区分民间“科学家”与科学家的最本质特征。许多民间“科学家”也常常用科学家在童年或少年时期的故事来为自己辩护。但被民间“科学家”们有意无意地予以忽略的是，他们的思维形式与真正的科学家在童年或少年时期的思维形式虽有一定的相似性，但这种相似性却永远地凝固在了那样一个年龄段上，仿佛自幼年起就停止了发育。民间“科学家”们虽然对

科学充满了雄心壮志，试图“研究”科学界最艰深、最宏大的课题，试图“推翻”科学界最有实验基础的理论，但他们数十年如一日的行为却只是在一个极低的水平上循环往复。他们可以花几十年的时间来做“研究”，却无意拿出几年的时间来系统地学习科学。科学界的文献是开放的，但由于他们无意接受系统的科学训练，从而在实质上放弃了阅读和理解科学文献的能力。因此他们的“理论”无论用什么时髦的科学术语来包装，用科学界的标准来衡量，都只是停留在一种十分原始的、伽利略之前的思维水准上。

这两条特征当然既不是完备的，也不是毫无例外的，想要在这样一个模糊的领域中建立一个绝对清晰的定义是一种徒劳。但这两条概括了绝大多数民间“科学家”的基本特征。

远离了系统的科学训练，远离了科学文献，民间“科学家”获取知识的主要来源是科普读物。因此大量民间“科学家”的出现也使我们看到了科普在向大众传播科学知识的过程中所显露出的一个薄弱环节：那就是科普对于现代科学的通俗化处理具有一定程度的误导性。

这么说让我自己觉得很不安，因为我非常敬重科普，希望这样的说法不会被理解为轻视或贬低科普。我想要通过本文表达的观点是，科普是好东西，但她所面向的读者群体决定了她有无可避免的局限性，她不能作为科学研究的完整背景。一个试图研究科学的人所需获取的基础知识绝不能止步于科普的层次。**科普的作用是让没有机会研究科学的人了解科学；让有机会研究科学的人喜欢科学，给他们一个“第一推动力”，让他们超越科普、接受系统的科学训练、继而投身于真正的科学研究。科普不应该起的作用是让有志于研究科学的人以为那就是科学，以为读过科普就算懂得了科学**。遗憾的是，科普对民间“科学家”所起的恰恰是它不应该起的作用。

科普在本质上是面向非专业读者的，因此对许多科学概念和理论——尤其是高度抽象的现代科学概念和理论——不得不做极大的简化。这其中最重

要的一个简化就是抽去了科学的数学框架，取而代之的是一些文字化的描述以及与日常经验的类比。与这种对科学概念和理论的简化相平行的，是对科学研究过程的简化。科学发现往往被简化成几个概念在科学家脑海里“灵机一动”式的组合。仿佛牛顿的万有引力定律真的就是被苹果砸了脑袋后“灵机一动”就想到了；仿佛爱因斯坦的广义相对论真的就是从几个像“升降梯实验”那样的理想实验中“灵机一动”就得到了。现代科学的研究既有灵感的显现，又有大量扎实而复杂的数学演算及实验，两者相辅而成。但在科普读物中前者给人留下的印象往往远远深于后者，因为前者大体上是概念之旅，既新奇浪漫又富有戏剧性，而后者相形之下不仅显得枯燥乏味，而且往往不是文字叙述所能够完全涵盖的。科普读物的这些局限性都极其明显地体现在民间“科学家”们的“理论”以及他们的“研究”方法上。

什么时候的科学是基本上没有数学结构的呢？那是古代的科学，比如我国古代的五行学说，古希腊的元素学说，等等。在那些学说诞生的年代里，概念和术语的简单组合、纯粹的思辨就可以成为科学（自然哲学）。但是自伽利略之后，科学逐渐脱离纯粹的思辨而进入了以实验和数学体系为主导的时代，现代科学因此而获得了令人赞叹的严密性和精确性。现代科学的这些特点在许多科普读物中都得到了强调，有时甚至是反复的强调。许多科普读物的作者本身就是第一流的科学家，他们深知科学的真谛，他们的科普作品中绝没有忽略现代科学的任何一个重要特征（因此我们讨论的是科普的“局限性”而非“缺陷”）。但现实的情况却是，同样的一部作品对读者所起的作用是和读者本身的知识背景密切相关的。接受过系统科学训练的读者（包括有学术基础的科普作者本人）会自然而然地将科普中的文字叙述与自己在科学训练或研究中的知识及经验相结合，从而获得完整而深入的理解；但对于没有接受过系统科学训练的读者来说，文字化的叙述往往就只会产生文字化的理解。这种理解对于普通读者来说是足够了，但对于一个有志于从事科学研究的人来说却是远远不够的。读100遍“爱因斯坦花了整整 $N$ 年才完成广义相对论”的故事，也远远不如自己动手花 $N$ 个小时来再现一遍爱因斯坦对水星近日点进

动值的计算更能体会科学研究的感觉，更能体会现代科学描述自然的方式。这就好比是一个学编程的人，看几本编程的书，却一行程序都不写是学不到编程的精髓的。这种“动手体会”的要求当然不是针对普通读者的，如果是的话也就不需要科普了。但是对于真正有志于从事科学研究的人来说却是必须的。

热衷于砍杀相对论的民间“科学家”们，在挥舞屠刀之前，可否先与现代科学的数学体系做哪怕只是一次这样的“亲密接触”？可否先对人类智慧几百年来的成就做哪怕只是一个细节上的深度了解？

一部分民间“科学家”之所以用自己浅陋不堪的“理论”去挑战现代科学，还往往能挑战得神气十足、老气横秋，乃至盛气凌人，其中很重要的一点就是他们是彻底地“轻装上阵”，他们不仅扔掉了现代科学的数学框架，也扔掉了现代科学背后庞大的实验基础。所以他们可以声称自己的一个没有任何定量结果，没有任何精密实验支持的“理论”超越或推翻了一个有坚实实验基础的科学理论。连科学是人类描述自然的一种努力——从而必须尊重实验观测——这样基本的原则都可以视而不见，现代科学在他们手中自然就变得可以任意宰割了。但是离开了这两者(数学框架和实验基础)，科学就退回到了伽利略之前的时代，这事实上也就是绝大多数民间“科学家”所能达到的最高水准(甚至连这样的水准也已经是一种高估，因为哪怕在伽利略之前也已经有不少的学者，比如哥白尼、托勒密等，用相当观测化和数学化的方式来构筑理论了)。民间“科学家”们如果意识不到科普以及他们建立在科普之上的知识体系的局限性，只怕永远也超越不了这一水准。

提出了“统一场论”的民间“科学家”们，可否告诉我们，原则上——也就是不劳您亲自动手，哪怕给个思路也行——如何用你们的理论来推算一个像水星近日点进动值那样的实验结果？

科普并无过错，不仅无过，且有大功。但科普有其局限性。这种局限性只有当她被有志于从事科学研究的人视为科学本身，并以之作为自己“研究”科学的基础时才会显现出来。记得小时候读过一则古老的哲学故事，说有一群

人居住在山洞里，面向石壁、背朝洞口。在日月星光的更替中，他们可以看到外部世界在石壁上的投影，于是他们研究起了投影的运动，日复一日，年复一年。但他们谁也没有转过身去看一眼山洞外的世界，他们一直以为那些投影就是整个的世界。科普就好比是那些石壁上的投影，她是科学的一组影像，而民间“科学家”们则好比是山洞里的那群人，他们在研究影像。

影像没有错，但它有局限性，研究影像也没有错，但如果认为影像就是整个的世界，那就错了。

2003 年 7 月 2 日写于纽约

# 什么是民间“科学家”

*One of the symptoms of an approaching nervous breakdown is the belief that one's work is terribly important.*

*Bertrand Russell*, 1930

## 一、新民科引发的问题

2003 年 7 月，我曾写过一篇有关民间“科学家”（简称民科）的文章：《从民间“科学家”看科普的局限性》[①]。在那篇旧作中，我归纳了民间“科学家”的两条主要特征：

(1) 民间“科学家”没有接受过系统的科学训练。

(2) 民间“科学家”无意接受系统的科学训练。

那篇旧作由于发表较早，在我的同类文章中影响较大，被包括维基百科

① 已收录于本书。

"民科"词条在内的很多网站引用。不过自那篇旧作发表以来，我逐渐意识到它所归纳的民科特性过于狭窄，只适用于早年常见的传统民科。这些年来，我接触到了很多新类型的民科，他们与传统民科有一个很大的区别，那就是带有"教授"、"研究员"、"博导"等学术头衔。当然，在腐败大潮席卷神州的今天，那些头衔不一定都货真价实（确切地说"价"可能是实的，但"货"不一定真）。但不可否认的是，也确实有一些民科是或者曾经是——以后者居多——货真价实的"教授"、"研究员"、"博导"等。那些人都曾受过系统的科学训练，从而并不符合那篇旧作所归纳的民科特征。但那些人的所作所为却与传统民科并无二致，即通过非学术渠道发布不被学术界接受的"论文"，宣称自己破解了重大科学难题，或推翻了重大科学理论①。

那些新民科的涌现，使我有必要重新讨论这样一个问题：什么是民间"科学家"？

## 二、有关民科的几个较具误导性或典型性的观点

我之所以要讨论这个问题，除了想弥补旧作的不足外，还有一个用意是想借讨论这个问题之机，顺便澄清一些有关民科的较具误导性或典型性的观点。那些观点大都来自过去这些年我接触到的民科及其同情者。若无意外，我希望本文成为我最后一篇有关民科的独自成篇之作（因为多写此类文章并无太大价值，反而会让我"日进斗敌"）。对过去接触到的与民科有关的较具误导性或典型性的观点一并做些分析，可以避免留下太多有可能使我旧话重提的由头。具体地说，本文将分析以下三种较具误导性或典型性的观点：

（1）"泛民科"观点。持这种观点的人认为民科这个概念是相对的，将别

① 需要说明的是，那些曾经受过系统科学训练的民科在行为模式上与传统民科还是有一定区别的，主要体现在他们不像传统民科那样"无知者无畏"，他们文章的措辞要比传统民科来得谨慎，语气不像后者那样斩钉截铁。

人视为民科的人（比如在下），在更高水平的人（比如诺贝尔奖得主）面前，自己也将被归为民科。这种"泛民科"观点的威力是巨大的，它让我想起很多年前看过的一部美国喜剧系列片《火星叔叔马丁》（*My Favorite Martian*）。在那部系列片中，火星人"马丁叔叔"的头顶可以升出一对具有隐形功效的天线，但有一次那天线出了故障，升起之后没能使"马丁叔叔"隐形。这下麻烦大了，因为那会暴露"马丁叔叔"的火星人身份。怎么办呢？"马丁叔叔"想出了一个高招，那就是让那种天线成为流行饰品。一旦大家都戴上那样的天线，"马丁叔叔"的天线就不再扎眼了。这种让所有人都变得相似的方法成为了保护"马丁叔叔"的最佳方法，用"马丁叔叔"自己的话说（大意）：把一棵树藏起来的最好办法就是把它藏在树林里。"泛民科"观点对民科所起的作用也是如此，它通过让所有人都变成程度不同的民科，而让真正的民科得以遁形（当然，这或许只是民科同情者们的一厢情愿，民科自己恐怕非但不想遁形，反而急切地想要展示自己独有的"天线"）。

（2）将民科与学术界的非主流研究相提并论的观点。众所周知，学术界的研究有许多类型，其中既有主流，也有非主流。非主流研究的存在对学术界是有价值的，不仅因为它们中的某些或许有朝一日会变成主流，或具有部分价值，而且也因为它们与主流研究的竞争有时能帮助揭示主流研究的不足之处，或促使主流研究者将自己的理论表述得更严密。但非主流研究按定义就意味着职位及同路人较少，从而在谋职、发表等方面带有一定的弱势性，这一点往往被民科引为同类。将民科与学术界的非主流研究相提并论，可以起到模糊民科与学术界界限的作用，从而间接提高民科群体的地位。

（3）对"民科"中的"民"字作字面解读的观点。持这种观点的人认为所谓民科，就是栖身"民间"的科学家，而学术界则是所谓的"官科"（因为"官"与"民"相对）。如果说将民科与学术界的非主流研究相提并论可以起到模糊民科与学术界界限的作用，那么将学术界视为与"民科"相对的"官科"所起的作用则恰好相反，那就是使民科与学术界划清界限，并对后者进行抹黑。因为在中国，"官"字所代表的形象是相当负面的，如果学术界跟"官"是一丘之貉，那

么很多人也许会出于对“官”的反感而宁愿支持“民”科。除此之外，将学术界视为“官科”还有一个好处，那就是便于民科用阴谋论的手法为自己的受迫害情结寻找依据，即把自己打扮成被“官”欺压的“民”，把自己的观点不被学术界接受说成是自己的创见被后者所打压（他们显然没有意识到，科学史上有无数比他们新颖百倍的创见都被学术界接受了）。

## 三、民科的定义

以上三种观点，是我这些年接触到的有关民科的观点中较具误导性或典型性的。要想澄清这些观点，有必要对民科这一概念做一个适当的定义。这个定义的思路在本文开头其实已经涉及到了，那就是从民科的行为及发布渠道入手。当我们把某些带有学术头衔的人列为民科时，所依据的正是他们的行为及发布渠道与传统民科相同。由此可见，从这一角度入手定义民科要比我那篇旧作所列举的背景特征更具适用性。不仅如此，从这一角度入手也比列举民科的其他特征，比如狂妄、偏执等，更具适用性，因为后者往往与民科的具体个性有关，不易一概而论，而且那些特征大都具有贬义，容易引起不必要的意气之争。有鉴于此，本文拟从行为及发布渠道入手，引进以下定义：

所谓“民间科学家”（简称民科），是指以非学术渠道为主，宣称推翻重大科学理论，或破解重大科学难题的成年人。

在应用这个定义前，让我们对定义中的若干用语作一些简短说明：

- “非学术渠道”是指除学术刊物、学术机构预印本、学术会议等正规学术成果发布渠道以外的其他渠道。其中目前最受民科青睐的是博客、论坛、垃圾邮件等渠道①。

① 这里用“垃圾邮件”一词，并非刻意贬低，因为“垃圾邮件”是指未经对方许可强行发到用户邮箱中的邮件（unsolicited mails），尤其是指同时发给多个用户的邮件（unsolicited bulk mails）。民科以邮件方式向别人发送“论文”时所发的往往正是符合此定义的邮件。

- “以非学术渠道为主”中的“为主”二字，是考虑到托学术腐败的福或单凭运气，民科们有时也能在学术刊物上发布“论文”，从而不宜一刀切。不过由于能被民科渗透的刊物通常水平较低，加上民科“论文”的水平更低，发表之后势必石沉大海，难以彰显“鸿鹄之志”，因此民科不管“论文”发表与否，都会以非学术渠道为主进行长期推销，以扩大影响①。这“为主”二字的另一个作用，则是防止有人以某些科学家也撰写博客或参与论坛活动为由，来混淆其与民科的区别。对后者来说，撰写博客或参与论坛活动并非发布论文、谋求承认的主要渠道。此外还要说明的是，这“为主”二字因涉及不同渠道间的比较，有时需要一定的时间才能做出可靠的判断（一般来说，民科通过非学术渠道对自己“论文”所做的推销越卖力，就越便于人们作出可靠判断）。
- 本定义所说的“科学”既包括自然科学（物理、天文等），也包括数学。
- 本定义所说的“重大科学理论”既包括意义或影响重大的理论（比如相对论、量子力学等），也包括其他具有坚实基础——从而往往能“牵一发动全身”——的命题、定理等（比如“尺规化圆为方的不可能性”等）。
- 本定义所说的“重大科学难题”既包括未解决的难题（如哥德巴赫猜想、黎曼猜想等），也包括已解决的难题（如四色定理、费马大定理等），因为重新“破解”后者也是民科们所热衷的。
- “成年人”三个字的使用，是为了避免将尚在系统求学阶段的年轻人列为民科。如我在旧作中所说，民科的某些特征与童年或少年时期的科学家有一定的相似之处，民科们时常利用这一点为自己辩护。一个合

① 顺便说一下，这一行为隐含着民科的“论文”无论发表与否，都未被学术界真正接受，以及民科对自己“成就”进行反复宣称等未在定义中直接列出的特点。

理的民科定义则必须将这种混淆排除在外①。

## 四、民科定义的应用

定义既已给出，我们就可以用它来分析一些东西了。

首先可以看到的是，上述定义与我旧作中所归纳的传统民科的两条特征是相容的（但涵盖面更广，因为它还涵盖了本文开头所提到的带有学术头衔的民科）。因为满足那两条特征的传统民科显然无法跻身学术界，从而必然只能以非学术渠道为主来宣布自己的“发现”。这表明传统民科符合上述定义。其次我们还可以看到，民科的若干常见言论与上述定义也有很好的相容性，甚至有一定的因果传承关系。比如正因为民科是以非学术渠道为主宣布自己的重大“发现”，从而往往要面对如此重大的“发现”为何要用如此“简朴”的渠道发布的问题，对此的“最佳回答”莫过于是把自己比喻成当代的哥白尼、布鲁诺，把学术界比喻成当年的教廷（或当今神州的官场），这正是民科言论中很常见的类型。而一些民科言论所展现出的病态的狂妄与偏执，则与自以为作出重大“发现”后成名欲的爆棚，及在学术渠道前“小扣（或猛踢）柴扉久不开”后的愤恨不无关系。

接下来让我们再用上述定义来分析一下前面提到的那几种具有误导性或典型性的观点：

（1）“泛民科”观点。这种观点的谬误之处在于忽略了上述定义中的“以非学术渠道为主”及“破解重大科学难题”、“推翻重大科学理论”等界定。**一个人是否是民科并不单纯取决于水平高低，即便要论水平，也应该论相对于自己研究目标而言的水平**。一个有一定水平的人若从事的是自己水平不能及的

① 另外可以补充的是，这里的“成年人”一词只是简略说法，并不等同于年龄意义上的成年人，由于它的作用是避免将尚在系统求学阶段的年轻人列为民科，因此其含义也是以是否仍处于系统求学阶段为界定的。一个年龄意义上的未成年人若在从事本定义所述的民科行为的同时，已不再接受系统的科学训练，那对于本定义来说就可被列为“成年人”。

“研究”(比如“破解重大科学难题”或“推翻重大科学理论”)而至偏执的程度(即无法以学术渠道为主进行发布却仍执迷不悟),他就会成为民科；而一位中学物理教师如果从事的是自己的教学研究,他就不是民科[①]。

(2)将民科与学术界的非主流研究相提并论的观点。这种观点的谬误之处在于忽略了上述定义中的“以非学术渠道为主”这一界定。**学术界的非主流研究与主流研究一样,都是以学术渠道为主发布成果的**。一旦离开那样的渠道,它们就不再是学术界的非主流研究了。只有在那时,它们才会与民科有可比性(可惜那时它们对提升民科群体的地位往往已不起作用了)。

(3)对“民科”中的“民”字作字面解读的观点。这种将民科理解为栖身“民间”的科学家,将学术界定义为“官科”的观点同样不符合上述定义。因为上述定义丝毫未涉及人在哪里的问题,它所关注的只是行为及发布渠道。一个身在民间的研究者如果以学术渠道为主发布研究成果,接受同行评议,他就不是民科(一个最典型的例子就是常被民科们引为“知己”的尚在专利局时的爱因斯坦)；反过来,一个身在学术界甚至有过杰出成就的人若只能以非学术渠道为主来宣称重大“研究”,那么无论他身在何处,名声是否显赫,起码在该项“研究”中的表现可被视为民科(带有学术头衔的民科就属于此类)。如果一定要对民科中的“民”字作一个字面解读的话,**那么虽然绝大多数民科确实身在民间,这个字的本质含义却应该界定为发布渠道的民间性**。

在本文最后有必要指出的是,如我在旧作中曾经说过的,对民科这样一个概念做任何定义或归纳都不可能做到完备或精确。本定义也不例外,除有可能存在反例或难以判别的个例外,其涵盖面也还不够广(虽比旧作来得广,却仍不足以涵盖全体)。比如由于将发布渠道作为定义的一部分,使得正在从事“研究”,但尚未发布任何消息(从而与外部社会尚处于绝缘状态)的人无论其“研究”多么民科化,都不在本定义的涵盖范围之内；又比如由于将“宣称推翻

---

① 打个比方来说：小蛇虽小,若吃的是小动物,那就是正常行为；大蛇虽大,若意在吞象,且不死不休,那就是民科行为。

重大科学理论，或破解重大科学难题”作为定义的一部分，使得“胃口”小，不以之为目标的人无论其“研究”多么民科化，也并不在本定义的涵盖范围之内①。

2011年3月5日写于纽约

2014年1月9日最新修订

① 不过，我见过的民科不少，那样的人却尚未见过，这或许并非偶然，而是因为“胃口”小，甘心做小课题，不好高骛远的人不容易成为民科。

# 学物理能做什么?①

说实话,接到这篇让我向年轻人介绍“学物理能做什么?”的约稿时,我的第一反应是婉拒。当然不是怕“年轻人”三个字把自己衬老了,而是觉得以我已经转行了的身份来写这样的文章,恐怕会适得其反。因为这篇约稿的背景,是物理在高考志愿中逐渐受到冷落,而约稿的目的,则是要鼓励年轻人选择物理。对于这个目的来说,我恐怕是一个坏榜样。不过约稿编辑洞察先机,在约稿信中直接把我归为“工作转行,却并没有真正离开物理”这样一类人的代表,断了我的托词。于是我只好老老实实来写这篇文章。

我体会编辑让包括我在内已经转行的人也来写这个话题,是想让年轻人知道,即便他们今后实际从事的是别的职业,也依然可以报考物理专业。因为他们在这一专业所受的训练,对从事别的职业同样会有助益,甚至会有独特的优势。这样的意思我在以前的文章中曾经作为体会述及过,但从未当作一种专业选择的策略向任何人推荐过,因为在我看来,物理所具有的这种优势是不能当作策略来用的。任何人如果出于喜爱物理以外的其他动机而选择物理,

① 本文曾发表于《现代物理知识》2010 年第 3 期(中国科学院高能物理研究所)。

其结果很可能是既学不好物理，也无法实现原本希望通过物理来实现的其他目标。因为物理对于不喜爱她的人来说，并不是一门容易的专业。

但另一方面，物理在高考志愿中所受的冷落，未必是因为越来越多的年轻人已不再喜爱物理，而很可能只是因为年轻人变得更现实了，或受到了来自亲朋好友更现实的劝告。我想本文的真正读者应该是这部分年轻人，而本文所要表述的观点是：请不要因为担心未来的出路而放弃自己喜爱的物理。这并非是劝诫任何人为了理想放弃现实，而只是说，起码就物理而言，这两者之间的距离并不像许多人以为的那样遥远，从而**没有必要**担心，更没有必要因此而早早地放弃自己的理想。一个人源自年轻时代的激情，在未来的人生之路上往往是难以再现的，给自己一个机会去追求并真正了解自己的兴趣，是明智而无悔的选择，过早地放弃——尤其是建立在错误理由之上的放弃——则是令人惋惜的。

好了，现在我们言归正传，从求职的角度来说说“学物理能做什么?”。其实这个问题基本上是不需要回答的，因为相反的问题——即学物理不能做什么——恐怕反而是比较困难的。在学物理所能做的事情当中，除了物理本身以外，还涉及许许多多其他职业，本文只举其中一个例子：金融。之所以举这个例子，除了金融是一种热门职业外，更重要的是因为这个曾经与物理风马牛不相及的职业，比其他职业更能体现出人们从学物理中获得的能力所具有的广泛适用性。

如果不考虑零星的个例，物理学家进入金融界大致可以追溯到 20 世纪 70 年代末的美国。当时由苏联发射人造卫星在美国引发的科技震荡及热潮已渐渐消退，很多物理专业的学生开始寻找新的求职领域。而在那之前不久，金融领域本身发生的一些变化，恰好为物理学家的进入创造了条件。1973 年，当时在芝加哥大学(University of Chicago)和麻省理工学院(MIT)的经济学家布莱克(Fischer Black，1938—1995 年)、斯科尔斯(Myron Scholes，1941— )及默顿(Robert C. Merton，1944— )等人提出了有关金融衍生品(Derivatives)的数学模型。这个数学模型(称为布莱克-斯科尔斯模型)的基

础是一组偏微分方程,而这组偏微分方程与物理学上用来模拟随机过程的某些方程式具有一定的相似性。显然,物理学家们在研究这种方程式上具有很大的优势。而且这种优势不仅仅来自于那些方程式与物理方程式之间的相似性,更多地是来自物理学家们所具有的处理包括那种方程式在内的各种复杂问题的普遍技巧,以及修正旧模型、构建新模型的能力。在瞬息万变的金融世界里,这种能力无疑具有极大的重要性。

金融衍生品在 20 世纪 70 年代时还是一种不太重要的东西,默顿在当年论文的开头甚至表示,为此发展一套理论也许是不值得的。但在 30 多年后的今天,金融衍生品的市场规模却远远超过了像股票那样的传统金融产品。1997 年,默顿和斯科尔斯因为当年那“也许是不值得的”工作获得了诺贝尔经济学奖(布莱克很遗憾地因为已经去世,无法分享这一荣誉)。而物理学家参与其中共同打造的这种以金融模型分析为主要职责的新角色,也早已成为了金融界的一种重要的新兴职业:定量分析师(quantitative analyst,简称 quant)。由于这一职业的兴起,在 20 世纪 90 年代,华尔街成为了向物理学家提供职位最多的领域之一。在某些公司中,物理学博士的人数竟然占到了公司总人数的三分之一甚至更多。到了 2007 年,就连物理学界最著名的论文预印本档案馆 arXiv. org 也为参与金融分析的物理学家们增添了一个新的论文类别:定量金融(quantitative finance)。这个类别如今每个月都有几十篇论文。

在物理学家眼里,一个领域的成熟往往意味着它的淡出。对于金融分析来说,这一天即便存在也还很遥远。事实上,富有戏剧性的是,就在默顿和斯科尔斯获得诺贝尔奖的第二年,这两人曾亲自出任董事会成员的著名对冲基金:美国长期资本管理公司(Long-Term Capital Management)就陷入了重大危机,被其他公司接管。而全球金融危机的爆发更是使很多人对金融世界究竟存不存在规律产生了怀疑。有人认为,早在 20 世纪 60 年代末,有“分形之父”美誉的数学家曼德布洛特(Benoît B. Mandelbrot, 1924— )就已经提出过,金融世界在本质上是混沌的。但另一些人则认为,即便金融世界果真是混

沌的，那也只不过是说我们无法进行长期预测，定量分析师仍然有可能通过分析短期规律来获取利润。究竟哪种观点正确，恐怕还有待于更多的讨论，这其中也不乏物理学家参与的余地。

在约稿信中，编辑曾建议我结合自己的经历谈谈体会，不过我想这对年轻人恐怕不会有劝导力，因为我并不是什么成功人士。我唯一能说的，是当我离开物理去做别的职业后，从未遇到过技术性的困难，所有的问题与我曾经解决或试图解决过的物理问题比起来，都显得相对简单。有时我会想到一个或许不太贴切的比喻：小时候我像很多其他小朋友一样，看电影《少林寺》入了迷，幻想着自己也能练一些武功，比如轻功。于是我让妈妈给我做了一对可以绑在腿上（但不能让小朋友们看出来）的沙袋，天天扎着走，期待有朝一日去掉绑腿后就算不能飞檐走壁，起码也能健步如飞。我觉得，学物理所受的训练就好比是扎着绑腿走路的那种锻炼，而转到别的职业后的感觉是去掉了绑腿。走路本身的难度并没有改变，但因为有了扎绑腿练就的基础，走路时就可能会觉得比较轻松。

我记得很多年前，人们曾经很看重学历，后来的一个鲜明转变是越来越多的人意识到了能力重于学历。类似地，专业曾经是很重要的求职凭据，但在日益注重能力的时代里，专业与职业的关联也在很大程度上让位给了能力与职业的关联。一个专业对口的人虽然能比其他人更快地投入工作，但这个优势往往只体现在最初的一小段时间里。一旦大家都熟悉了业务之后，究竟谁更有效率，谁更能处理复杂问题，谁更能应对尖锐挑战，终究还是要看能力。而学物理对能力的训练是比较全面的，既有严密的数学和逻辑，又能与现实数据打交道，这个专业具有广泛的适用性是不足为奇的。

在本文的最后，请允许我再强调一次：本文的目的不是鼓励不喜爱物理的人通过学物理来达到其他目的（比方说，如果你想做的原本就是金融，那就不要去学物理），而只是想告诉喜爱物理的年轻人，学物理不是单行道，不要为出路担心，更不要因为无谓的担心而过早地放弃物理。美国物理学家费恩曼（Richard Feynman，1918—1988 年）在去世前不久曾收到过一位父亲的来

信,为自己即将进大学的孩子的前途问题征询意见。费恩曼在回信中提了这样一条建议:“别考虑你想成为什么,只考虑你想做什么。”

喜爱物理的年轻朋友,如果你现在想做的是学物理,那就听费恩曼的话,大胆地去做吧。

2010 年 4 月 22 日写于纽约

# 关于普通科普与专业科普

本文的主要目的是叙述一下我对科普——尤其是数学、物理类科普——的某些零星想法，作为对拙作《黎曼猜想漫谈》的后记所提到的“普通科普”与“专业科普”这两个概念的注释，并对专业科普的价值略作评述。

我觉得普通科普（即基本不用数学公式的科普）比较适合于介绍那些容易进行通俗类比的东西，因为通俗类比是向普通读者介绍技术性内容的最有效的手段之一，通常具有将定量转化为定性，将不熟悉概念转化为熟悉概念的作用（当然，往往会因转化而导致拙作《从民间“科学家”看科普的局限性》所述的那些局限性）。对于不容易进行通俗类比的内容，普通科普则会面临不小的困难，并且常常会陷入这样的困境：即对某些无法回避的数学公式或技术性内容不得不进行缺乏类比，或类比得不太贴切的文字描述，有时甚至不得不对数学公式进行文字化的“直译”或复述——后者或许可以称为“文字公式”。

“文字公式”相较于数学公式来说，其实往往是更不容易理解的东西。事实上，从历史上讲，数学符号之所以被引入科学，乃是因为它有着文字无法替代的简单性和清晰性。从这个意义讲，**从文字到数学符号乃是往简单和易于理解的方向迈出的一步，而不是相反；而对数学公式进行文字“直译”或复述，**

**反倒是在一定程度上重新退回到了“史前”科学的繁琐、晦涩及模糊**。那样的叙述虽然在表面上避免了被科普界视为“票房毒药”的数学公式，给人以普及的印象，实际上却未必比直接使用数学公式更具普及性。因为没有数学基础的读者读到这种“文字公式”后，虽然每个字都认识，却未必能把握整句话的确切含义(或产生一个把握了的错觉)；而有一定基础的读者看了这种“文字公式”则可能会有隔膜感，会在脑子里试图将“文字公式”还原成数学公式，却远不如直接看到后者来得轻松透彻。因此，对于那种为回避数学公式而不得不诉诸“文字公式”的题材，使用“文字公式”的实际结果有可能是两头不讨好，即既不能有效地帮助普通读者理解公式的含义，也投不了有一定基础的读者所好。对于那样的题材，我觉得专业科普(即介于普通科普与专著之间、不回避数学公式的科普，有时也称为“高级科普”，不过我更倾向于“专业科普”这一术语，以避免因“高级”一词造成普通科普“低级”的不必要的攀比印象)有很大的施展余地。

当然，对数学公式与文字的难易评判不可一概而论，复杂到一定程度的数学公式自然绝非普通读者所能理解。比如黎曼-西格尔(Riemann-Siegel)公式就是一个例子①。但即便那样的公式，也有某些特殊的价值，比如在向读者介绍计算黎曼 $\zeta$ 函数非平凡零点的难度时，我们固然可以搜肠刮肚地找出一系列形容词来加以描述，或者用数学家们计算零点的艰辛努力来作间接说明，但让读者亲眼看一看黎曼-西格尔公式的复杂性，也不失为是一种方法。哪怕看不懂，只当插图来看，也有可能起到一种更直接，甚至印象更深刻的说明作用。

另一方面，即便对于普通科普能够胜任的内容来说，过分排斥公式在我看来也是不必要的谨慎，甚至可以说是某种程度上的误区。对这一误区最直白的描述也许是英国物理学家霍金(Stephen Hawking)在某一版的《时间简史》(*A Brief History of Time*)的前言或后记中引述过的一句编辑的警告：每一

① 黎曼-西格尔公式是一个计算黎曼 $\zeta$ 函数非平凡零点的复杂公式，具体形式可参阅拙作《黎曼猜想漫谈》(清华大学出版社，2012 年)第 11 章。

个数学公式都会使读者减半。记得霍金在引述了那句警告后，表示自己在书中还是用到了一个公式：$E=mc^2$。他并且风趣地表示，希望那不会使该书的读者减少一半。霍金的胆子算是比较大的，更多的科普作者恐怕宁肯用“能量等于质量乘以光速的平方”那样的“文字公式”来代替 $E=mc^2$ 这样的数学公式。但仔细想想，那样的“文字公式”果真比数学公式更容易普及吗？有多少读者是知道什么叫做“平方”，却不知道它在数学上是用右上角的“2”来表示的？更何况，“质量乘以光速的平方”中的“平方”究竟是指“光速”的平方，还是“质量乘以光速”的平方，在“文字公式”中是分不清的，而初中甚至高小水平的读者多半就已经知道像 $E=mc^2$ 那样的数学公式中的平方是 $c$ 的平方，而不是 $mc$ 的平方，因为后者会被写成 $E=(mc)^2$，而不是 $E=mc^2$。数学公式的明晰性在这么一个小小的例子中都能显现出来，普通科普却千方百计地试图避免，不能不说是某种程度上的误区。

回到黎曼猜想这一题材上来。以上所说绝不是暗示我所读过的那两本有关黎曼猜想的科普书已经陷入了那样的误区或困境①。事实上，对于黎曼猜想这样一个高度技术性的数学题材来说，那两本书在深入浅出方面所做的努力是很值得钦佩的，而且它们各自都使用了少量的数学公式。不过，读者看完那两本科普后，对数学故事毫无疑问会留有印象，但对黎曼猜想本身究竟能知道多少，或许仍是可疑的。因为在对数学公式作了较大幅度的回避之后，容易出现这样的情形：即数学故事中数学的面目远比人物的面目来得模糊。而数学故事中数学的面目一旦模糊了，那么故事的背后究竟是黎曼猜想、费马猜想、还是哥德巴赫猜想，也有可能会变得模糊起来。若干年之后，看过黎曼猜想、费马猜想，或哥德巴赫猜想科普书的读者或许只会记得这样的共同场景：那就是一群数学家作了各种各样的努力，经历过各种各样的趣事，试图解决一

---

① “那两本有关黎曼猜想的科普书”指的是拙作《黎曼猜想漫谈》的后记所提到的德比希尔(John Derbyshire)的 *Prime Obsession*：*Bernhard Riemann and the Greatest Unsolved Problem in Mathematics*(Joseph Henry Press，2003)和索托伊(Marcus du Sautoy)的 *The Music of the Primes*：*Searching to Solve the Greatest Mystery in Mathematics*(Harper，2003)。

个著名的数学猜想。但他们试图解决的是什么猜想,他们各自究竟做了什么?则有可能只是记忆中的一团迷雾。我觉得专业科普在驱散这团迷雾上也能有一定的作为,可以作为普通科普很好的补充。

以上是对《黎曼猜想漫谈》一书所采用的专业科普这一定位的一点说明。关于普通科普与专业科普这一话题本身,当然还有很多其他可以谈论的地方,绝非本文这样的零星叙述所能涵盖。

2011 年 3 月 9 日写于纽约

# 人名索引

**Q**



**S**



**T**



**W**

# 术 语 索 引

**D**



**E**



**F**



**G**

**M**



**N**



**O**



**P**



**Q**

**Y**



**Z**